ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

Stress Level Detection from Human Speech Using Machine Learning Techniques

Deepti Patil[1], Nikita Mhetre[2], Shweta More[3], Dr. Bageshree Pathak[4], Dr. Nitin Palan[5]

B.E Student, Department of E&TC, Cummins College of Engineering for Women, Pune, Maharashtra, India[1]

B.E Student, Department of E&TC, Cummins College of Engineering for Women, Pune, Maharashtra, India [2]

B.E Student, Department of E&TC, Cummins College of Engineering for Women, Pune, Maharashtra, India [3]

Professor, Department of E&TC, Cummins College of Engineering for Women, Pune, Maharashtra, India[4]

Professor, Department of E&TC, Cummins College of Engineering for Women, Pune, Maharashtra, India[5]

ABSTRACT: The paper introduces a method to detect presence of emotional stress in human speech. Stress is the most significant factor that affects the mental health of a person and urges him to end his life. The paper throws light on a possible method to detect this stress so that the person with a high stress level can be given special attention and counseling. This paper analyzes the variations that human voice undergoes with respect to various emotional states. The features considered for differentiation of emotions are MFCC (Mel-Frequency Cepstral Coefficients), BFCC (Bark Frequency Cepstral Coefficients), pitch, energy, power and standard deviation. The database used for this consists of 60 utterances of each emotion i.e. Anger, sad, joy and neutral by 4 female and 2 male speakers of similar age groups and has been recorded in Marathi language. The classifiers used for classifying the emotions is the K-Nearest Neighbor (K-NN) and the Support Vector Machine (SVM).

I.INTRODUCTION

Speech signals are the most effective means of conveying information about person's feelings and emotions. Therefore, the speech emotion recognition system proves to be really useful in applications that include human-machine interaction such as web movies, in-car board system, computer tutorial etc. where the response of these systems to its users alters depending upon the detected emotions.

The human race in today's world is continuously striving for betterment of life. People, right from the early days of schooling, grow up in a very competitive environment. In the process of proving one's mettle to survive in this competition, humans have to face many stressful situations. This stress builds up as the person enters the corporate world and has to manage corporate life along with family life. If this stress goes unnoticed and increases beyond a certain extent, it naturally urges the person to end his life. A recent survey carried out by the World Health Organization in 2016, has brought to light that around 800000 people die due to suicide every year and that suicides is the second leading cause of death among the 13-35 year olds. It therefore, becomes important to manage this stress at the right time so as to prevent kind of extreme situation. Solution to this is to detect the stress level in the speech of the person and if it exceeds a specified limit, then that person needs to be given special attention and counselling. In this paper, the various modifications that human speech undergoes while he/she is under stress, have been analyzed so as to detect the level of stress so as to save a life before situation goes out of control.

II. LITERATURE SURVEY

Table 1 shows the comparison between 5 papers based on the features extracted, classifier used, emotions involved, database used and the accuracy of model. Paper 1 uses MFCC as the feature extraction technique and SVM, DSVMas the classifiers with a testing accuracy of 63.48%.Paper 2 shows that the emotions- fear and neutral have been classified using the K-NN and SVM classifier along with decision tree based on pitch feature extracted. It gives an accuracy of

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

78.7%.Paper 3 shows a comparative study between 2 feature extraction techniques viz. MFCC and BFCC. The classifier used is the Hidden Markov Model (HMM). The results show that BFCC retrieves more information than MFCC.

Paper 4 shows that the K-NN classifier classifies 4 emotions viz. anger, joy, sorrow, neutral based on features such as pitch, energy, entropy and MFCC. Both, paper 2 and 4 use the EMO database for feature extraction. The accuracy obtained is 72.22%.Paper 5 uses the MFCC feature extraction along with energy from a Chinese database classified using neural networks to classify 4 emotions with an accuracy of 54.2%.

Table 1

Sr. No.	Title of paper	Emotions	Feature Extracted	Classifier	Database	Accuracy
1.	Emotion Recognition in Speech Using MFCC with SVM, DSVM and Auto-encoder (2018)	Anger, Disgust, Happy, Neutral, Sad, Surprise, Fear	MFCC	SVM, DSVM	SAVEE database	63.48%
2.	On the use of pitch-based features for fear emotion detection from speech (2018)	Fear, Neutral	Pitch	Decision tree, K-nearest neighbours, support vector machine and subspace discriminant analysis	The EMO Database	78.7%
3.	Comparison of Warping Filter Banks using MATLAB Based Feature Extraction Techniques in ASR (2016)	-	MFCC, BFCC	HMM	-	Result :- BFCC retrieves more information than MFCC
4.	A Feature Extraction Scheme Based on Enhanced Wavelet Coefficients for Speech Emotion Recognition (2015)	Angry, Happy, Sad, Neutral	Pitch, Energy, Entropy, MFCC	KNN	EMO-DB	72.22%
5.	Speech Emotion Recognition Based on Data Mining Technology (2015)	Angry, Happy, Surprise, Sad	MFCC, Energy	Neural network	Chinese Database	54.2%

III. DATABASE CREATION

The database or speech corpora has been created by 4 female and 2 male speakers in **Marathi** language and comprises of 4 emotions namely anger, sorrow, joy and neutral. 12 utterances for each emotion have been recorded by 6 individuals which sums up to a total of 288 utterances.

The recording has been done using the voice recorder android app in a silent environment. Three age groups have been involved while recording- <15 years of age, 20-30 years of age and 40-45 years of age. The sampling frequency used is 44.1 kHz with an average bitrate of about 87.66 bps.The recorded file format is 'mp3' which is a compressed file format for digital audio. The mp3, also called the MPEG layer III (Moving Picture Experts Group layer III) format is founded by Karlheinz Brandenburg in 1993.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

IV. METHODOLOGY

The 2 phases used for achieving the results include the training phase and the testing phase. The block diagrams for both are shown in figures 1 and 2 respectively:

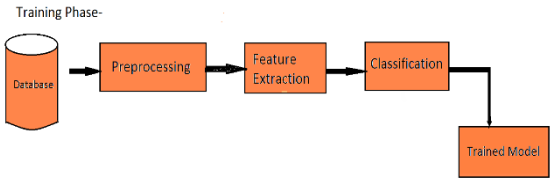

Figure 1

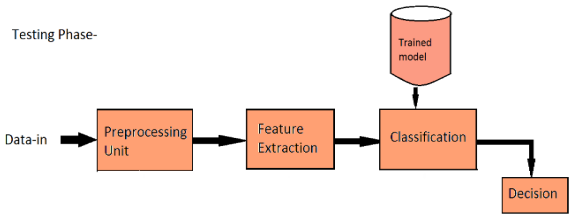

Figure 2

1. Signal pre-processing:

The pre-processing stage includes separation of voiced and unvoiced parts from the input speech signal. The algorithm works in such a manner that the silence and pauses in the signal are removed using the thresholding method. The algorithm used for silence removal is as follows:

i) Divide the signal into a finite number of frames.

ii)Traverse and analyze each frame and choose a suitable threshold value in terms of amplitude under which the signal is considered to be silent.

iii)Create an all-zero array of the size equal to that of input signal. Put only those values into the array that exceed the selected threshold. The resulting array returns the silence-removed speech signal shown in figure 3.

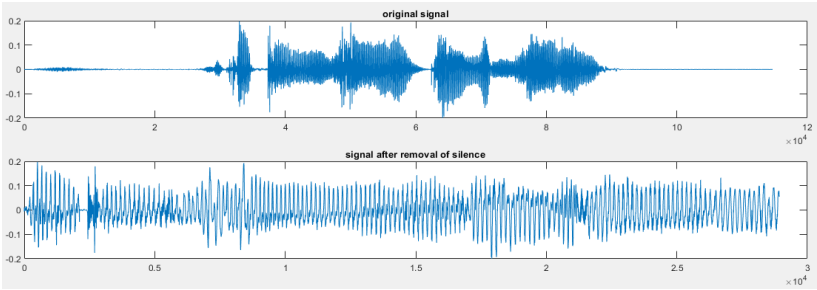

Figure 3

2. Feature Extraction:

The features that have been extracted from the database along with the algorithms used have been described below:

i.)Pitch

The pitch or the fundamental frequency f0, is defined as the relative highness or lowness of tone as perceived by the human ear.It basically depends on the number of vibrations produced by the vocal chords of a human per second. Here, the pitch is extracted using the normal correlation function. The pitch plot obtained is as shown in figure 4.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

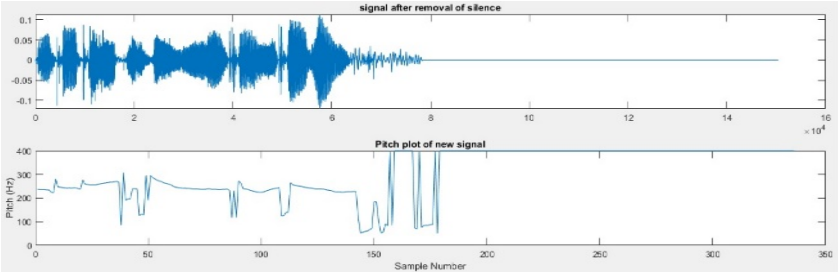

Figure 4

ii) Mel-Frequency Cepstral Coefficients

Mel- Frequency Cepstral Coefficients (MFCC) is a robust and dynamic method of feature extraction from speech signal. It is based on the known variation of human ear's critical bandwidth frequencies along with mel-filters that are spaced at low frequencies. The conversion from normal frequency to the mel frequency can be obtained by using the conversion formula:

$$\textbf{Mel (f) = 2595*log}_{10}\textbf{ (1 + f / 700)}$$

iii) Bark-Frequency Cepstral Coefficients

This is another effective method of extracting features from speech signal which produces better results than MFCC. This perception is expressed in a scale known as bark scale which shows a relationship which is non-linear with the sound wave frequency.

The algorithm remains the same as MFCC just with a difference in the conversion formula which is as stated below:

$$\textbf{B(f) = 6*log}_{10}\textbf{ [(f / 600)+((f / 600)}^{2}\textbf{ + 1))}^{1/2}$$

The algorithm to obtain BFCC and MFCC is given below in figure 5:

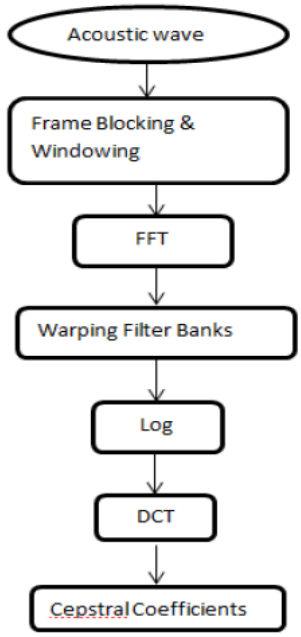

Figure 5

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

iv) Energy

Energy of a signal is also one of the features that varies with respect to the emotion content in speech. The paper determines the energy of the emotional speech by summing the squares of all the signal components. The formula stated below computes the energy content of the signal.

$$E = \sum x(n)^2$$

v) Mean

Mean of a speech signal equals to the amplitudes of the signal samples averaged over the total number of samples.

$$M = \sum x(n) / (\text{Total number of samples})$$

vi) Standard deviation

Standard deviation is a measure of how much the signal components deviate from the mean value of the signal. It quantifies the amount of variation or dispersion of the data samples. The mathematical formula for computing the standard deviation (SD) is shown below:

$$SD = (1 / N-1)* \sum (F-mean)^2$$

Where, N: total number of samples

F: individual signal component

3. Classification:

In this paper, the entire emotional speech corpora has been classified into 4 major emotions viz. Anger, joy, sorrow and neutral.

The classifiers chosen according to the proposed algorithm are the K Nearest Neighbor (K-NN) classifier and Support Vector Machine (SVM) classifier -which is a binary classifier. Both these classifiers are simple but equally strong means of classifying data into 2 or multiple classes;provided the data is labeled (Supervised learning).

Classification using K-NN:

K-Nearest Neighbor algorithm is the simplest of the classification algorithms which classifies 2 or more classes based on the case data using distance measures, specifically, the Euclidian Distance measure.

Consider 2 classes A and B. A random data point 'c' can be classified as either class A point or class B point based on the Euclidian distances calculated between the point 'c' and already classified points and selecting the minimum distance from them.

'K' denotes the number of minimum distances to be considered.

The Euclidian distance between any 2 points a(x, y) and c(x1, y1) is given by the mathematical equation,

$$\text{Distance} = [(x-x1)^2 + (y-y1)^2]^{1/2}$$

Classification using SVM:

The Support Vector Machine is a classifier which is discriminative in nature. It is formally defined by a separating hyperplane. In other words, given labeled training data, the algorithm gives output in form of an optimal hyperplane that is capable of categorizing new examples. Only the points known as 'support vectors' are significant. These are the points that are closest to the opposing class. The SVM classifier finds a hyperplane given by

$$g(x) = w^T x + b$$

that correctly separates 2 classes with maximum margin. Consider 2 classes A and B. SVM finds w and b such that function g(x) equals to 1 for nearest data points belonging to class A and -1 for the nearest ones of class B.

V.RESULTS

The 6 features that are extracted from the feature vectors which are given as an input to the classifier model.

The figures 6 to 11 show the feature plots obtained for the 4 emotions mentioned.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

Figure 6

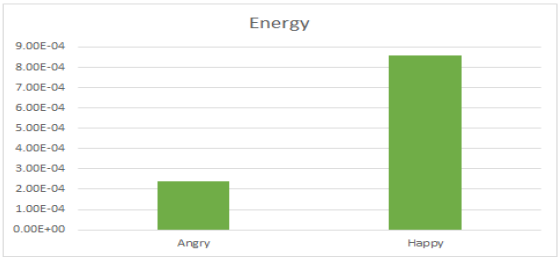

Figure 7

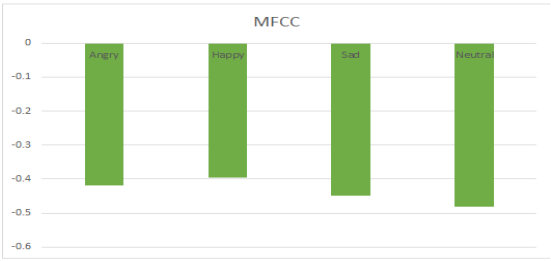

Figure 8

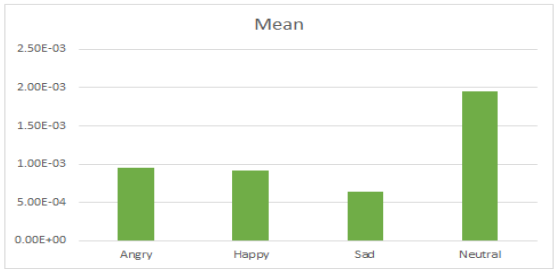

Figure 9

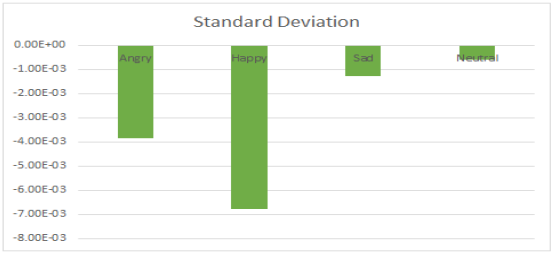

Figure 10

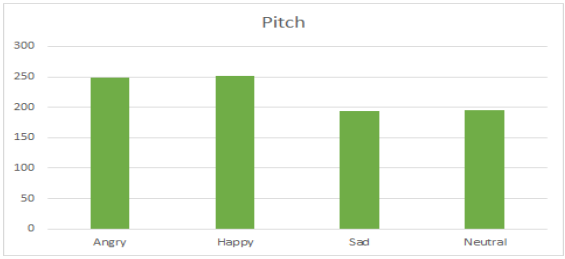

Figure 11

The classification accuracy obtained using the K-NN classifier is 66% and that obtained using SVM is around 70%.

VI. CONCLUSIONS

The feature extraction techniques used in this paper are MFCC and BFCC, energy, mean, standard deviation and pitch. Classification algorithms used are K-NN and SVM.

From the feature plots obtained we conclude that BFCC, Energy, Pitch, Mean and Standard Deviation are dominant contributors for differentiating among the 4 emotions. It is also observed that MFCC extraction technique does not produce desired results.After classification it is concluded that K-NN algorithm gives an accuracy of 66% where as SVM gives 70% accuracy.

REFERENCES

[1] Ying SHI, Weihua SONG, Speech Emotion Recognition Based on Data MiningTechnology, 2010 Sixth International Conference on Natural Computation (ICNC 2010), 2015.
[2] C. Shahnaz and S. Sultana,Bangladesh A Feature Extraction Scheme Based on Enhanced

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 8, Issue 3, March 2019

Wavelet Coefficients for Speech EmotionRecognition, 978-1-4799-4132-2/14/$31.00 ©2014 IEEE.
[3] Safa Chebbi and Sofia Ben Jebara ON THE USE OF PITCH-BASED FEATURES FOR FEAR EMOTION DETECTION FROM SPEECH 4th International Conference on Advanced Technologies For Signal and Image Processing ATSIP2018 March 21-24, 2018,Sousse, Tunisia.
[4] Hadhami Aouani, Yassine Ben Ayed Emotion Recognition in Speech Using MFCC with SVM, DSVM and Auto-encoder,4th International Conference on Advanced Technologies 2018
[5] S.Suganya, M.Selvaganapathy, N.Nishavithri Comparison of Warping Filter Banks using MATLAB Based Feature Extraction Techniques in ASR, DOI: 10.15680/IJIRCCE.2016. 0405018