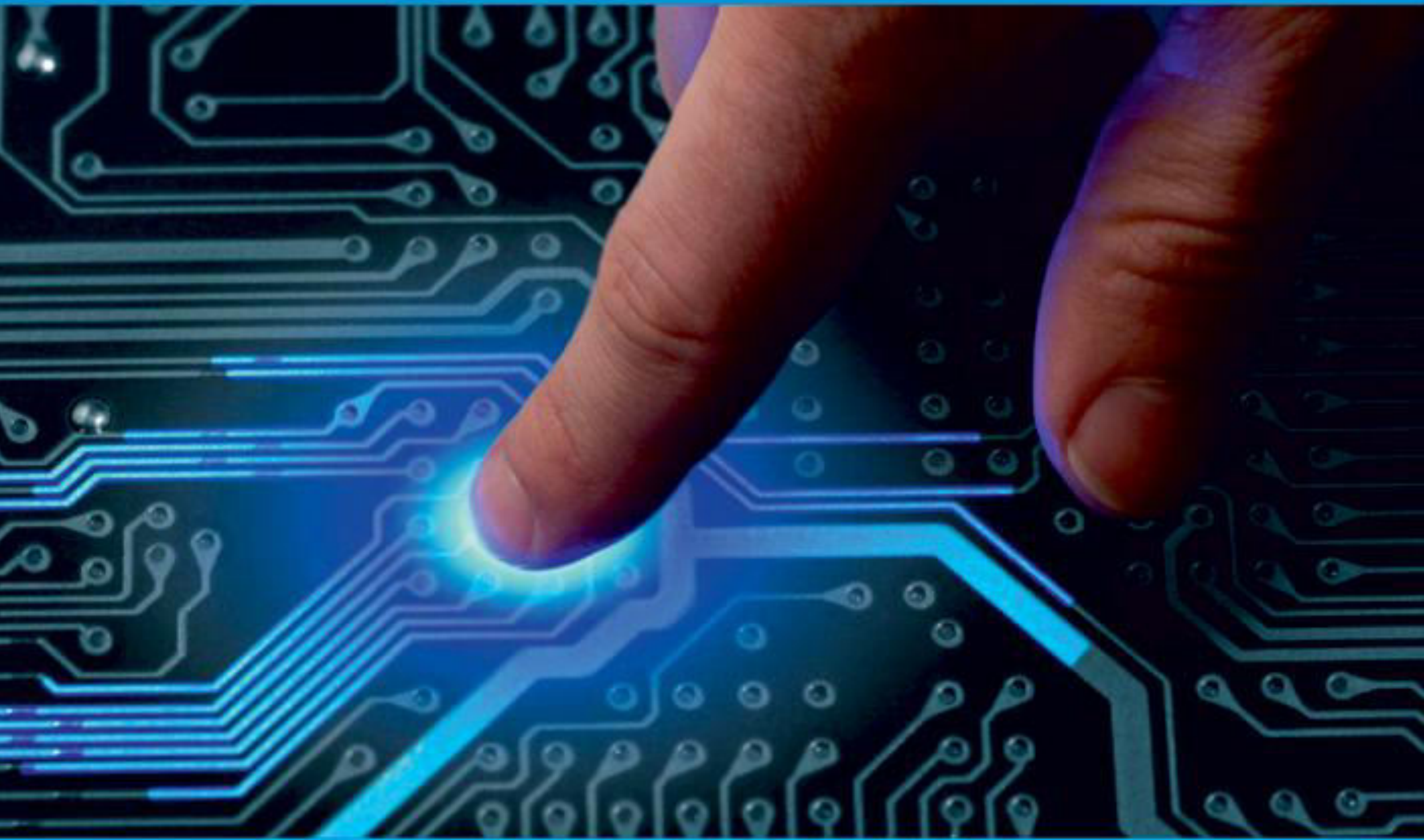
IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 9, Issue 7, July 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.542

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 7.542

|| Volume 9, Issue 7, July 2021 ||

| DOI: 10.15680/IJIRCCE.2021.0907140 |

Emotion Recognition in Speech with MLP Classifier and MFCC Feature

Mrunal Patil[1], Mayuri Kurkure[2], Gaurav Jumde[3], Saurabh Joshi[4],Vaibhav Patil[5], Ashish T. Bhole[6]

Under Graduate Student, Department of Computer Engineering, Shram Sadhana Bombay Trust's College of Engineering and Technology, Jalgaon, Maharashtra, India[1,2,3,4,5]

Associate Professor, Department of Computer Engineering, Shram Sadhana Bombay Trust's College of Engineering and Technology, Jalgaon, Maharashtra, India[6]

ABSTRACT: Speech is natural way of expressing ourselves. Many researchers today consider speech signal as quick and effective process to interconnect between computer and human therebymaking it a more challenging component of human computer interaction (HCI). Many techniques have been used to recognize emotions from speech. The proposed novel emotion recognition system uses Multi-Layer Perceptron (MLP) classifier and Mel-Frequency Cepstral Coefficients (MFCC)to classify speech signal fordetecting embedded emotions in human voice. The proposed system builds a learning model to analyse speech signal & prediction and define the accuracy of the model.

KEYWORDS: Speech, Emotion recognition, Machine learning, Neural Network (NN), Multi-Layer Perceptron (MLP),Mel-Frequency Cepstral Coefficients (MFCC)

I. INTRODUCTION

Speech is the fast and normal way of communicating between the humans. Considering a speech signal is an effective process for communication between humans and computer. It means computer should have knowledge to identify human voice. We define a system which recognizes emotions from speech. Although it is a challenging task because recognizing emotion are hard toidentify. But recognizing emotions from speech plays an important role in Human-Computer Interaction. This system is beneficial for robots, forensic science, medical field, customer services, etc. The system is collection of methodologies that process and classify a speech signal to detect emotions embedded in human voice.

The approach of recognizing emotions from speech mainly categorizing into three different phases. These are data preprocessing, feature extraction and classification. Accuracy of the model is depends on level of naturalness of the database which is used as an input to the system. For this project recorded datasets is used. It contains emotional speech and songs. Data require preprocessing before extracting the feature. Feature extraction is the process of identifying important attribute in the data. It aims to reduce the number of feature in a dataset by creating new feature from existing once [1]. Features are important for classifier. For example feature should be on energylevel of voice, voice quality, harsh, etc All features can passed to the classification process. Classification is the process of categorizing a data into classes. There are various classifiers are available for proper classification of data. Each dimension of system reflects psychological characteristics of the emotions.

II. RELATED WORK

Speech Emotion Recognition is implemented by executing the methods that include separation of the speech signal of audio file and extraction of features for the final recognition. In terms of acoustics, speech processing methods offer very valuable information about communication like volume, speed, intonation of a voice along with gestures and other non-verbalcues derived mainly from prosodic and spectral features. Both frameworks deal with a very challenging problem because emotional states have no clear boundaries and often differ from person to person [2]**.**Threesignificant view of designing a speech emotion recognition system. The first is to select appropriate features to represent the speech. The second problem is to design a suitable classificationscheme, and the third problem is to fully prepare the emotional speech database to evaluate the performance of the system[3].

From several years, emotion recognition is an important area for researchers. Reason behind this is to make intelligent system more efficient and accurate. Multiple research theory and models are available on emotion recognition from speech. But still there is accuracy is required [4]. In [5] author proposed a ranking SVM method for synthesize information about emotion recognition. This ranking method, train SVM algorithms for particular emotions, treating data from every sound as a distinct query then mixed all predictions from rankers to apply multi-class prediction. Ranking SVM attain two advantages, first for training and testing steps in speaker- independent it create speaker specific data. Secondly, it considers the percipient that each speaker may express mixed of emotion to recognize the commanding emotion. Ranking approaches attain substantial gain in terms of accuracy compare to conventional SVM in two public datasets of acted emotional speech, Berlin and LDC.

In both acted data and the spontaneous data, which comprises neutral intense emotional observation, ranking-based SVM achieved higher accuracy in recognizing emotional utterances than conventional SVM methods. Unweight average (UA) or Balance accuracy achieved 44.4%. In [6], author proposed a new system for emotion classification of observation signals. The system uses a Short time logs frequency power coefficients (LFPC) and discrete HMM to distinguish the speech signals and Classifier Respectively. This method categorized the emotion into six different variety then used the private dataset to train and test the new system. In order to judge the performance of the proposed method, LFPC is differentiate with the mel-frequency Cepstral coefficients (MFCC) and linear prediction Cepstral coefficients (LPCC). Outcome illustrate the average and best classification perfection achieved 78% and 96% respectively. Additionally, results shows that LFPC is a better option as feature for emotion classification than the standard features.

Goal to improve speech emotion recognition in speaker-independent with three level speech emotion recognition method. This method categories different emotions from start to end then select proper feature by using Fisher rate. The output of Fisher rate is an input parameters for multi- level SVM based classifier. Furthermore principal component analysis (PCA) and artificial neural network (ANN) are employed to reduce the dimensionality and classification of four comparative experiments, respectively. Four comparative experiments include Fisher + SVM, PCA + SVM, Fisher + ANN and PCA + ANN. Consequence denote dimension in reduction Fisher is better than PCA and for categories, SVM is more expansible to compare ANN for emotion recognition in speaker independent . The recognition rates for three level are 86.5%, 68.5% and 50.2% separately in Beihang university database of emotional speech (BHUDES) [7].

In [8], author presented a new modulation spectral features (MSFs) human speech emotion recognition. Some important features extracted from an auditory-inspired long-term spectro-temporal by utilizing a modulation filter bank and an auditory filter bank for speech decomposition. Important data which is missing from traditional short-term spectral features achieved from acoustic frequency and temporal modulation frequency components which is get from above method. For classification process, SVM with radial basis function (RBF) used. In experiments, MSFs gives better performance than Mel-Frequency Cepstral Coefficients (MFCC) and Perceptual Linear Prediction Coefficients (PLPC). When MSFs use enhanced prosodic features, it increases the recognition performance. Additionally, recognition rate of 91.6% is achieved for classification. In [9], identified a new spectral feature in order to determine emotions from speech and to classify groups. This study is mainly based on acoustic feature and novel hierarchical classifier, in which emotions are grouped. For classifying emotions there are different classifiers are used. Such as GMM, HMM and MLP classifiers are used. The innovation of the proposed method is based on two things that is selection of an important feature and second is extracted features are classified class-wise. For the proposed model Berlin dataset is used. As a system uses a novel hierarchical classifier, the hierarchical approach achieves better performance. Performance of HMM method achieves 68.57% accuracy and hierarchical model gives 71.75% accuracy.

In [10], author introduce an ensemble random forest to trees (ERF Trees) method with a large quantity of features for emotion recognition without referring any language or semantic information remains an unsolved problem. Ensemble random forest to trees (ERF Trees) method is applied on mini size of data with large quantity of features. In order to estimate the proposed method an experiment results on a Chinese emotional speech dataset designates, this method establish improvement on emotion recognition rate. Additionally, ERF Trees performs better than popular dimension reduction methods such as PCA and multi-dimensional scaling (MDS) and newly developed ISO Map. The greatest perfection with 16 features for female dataset achieved the highest accurate rate of 82.54%, while the worst is only 16% on 84 features with natural data set.

Fusion-based method for speech emotion recognition by employing multiple classifier and acoustic-prosodic (AP) features and semantic labels (SLs) proposed by author [11]. In this fusion method, first AP features are extracted then three different types of base-level classifier include GMMs, SVMs, MLP and Meta decision tree (MDT) are used. The maximum entropy model (MaxEnt) in the semantic labels method are applied. To define the emotion recognition outcome, In the final state the integrated consequence from the SL-based and AS-based are utilized. The experimental result on private dataset shows the performance based on MDT archives 80%, SL-based recognition archives 80.92, and mixture of AP and SL archives 83.55%.

In [12], author proposed speech emotion recognition theory from call center applications. For this research Narayanan focuses mainly on detecting negative and non-negative (anger and happy) emotions. Different types of information include acoustic, lexical, and discourse are used for emotion recognition. In addition, information-theoretic contents of emotional salience is presented to obtain data at emotion information at the language level. For this K-NN and Linear discriminate classifier are used for different types of features. Outcomes demonstrates by combining three information source instead of one source, classification accuracy increases by 40.7% for males and 36.4% for females. Compare to pervious work improvement range in accuracy is from 1.4% to 6.75% for male and 0.75% to 3.96% for female.

A novel set of features of speech emotions recognition represented by Yang & Lugger [13]. These features are mainly on music theory. First step is to predict pitch of signal and then computing spherical autocorrelation of pitch histogram. It calculate the incidence of dissimilar two pitch duration, which cause a harmonic or inharmonic impression. Bayesian classifier is used for classification. Berlin emotion database is used for proposed theory. Harmony features indicate an improvement in recognition performance. Average rate of recognition is 2%.

In [14], author represent a hierarchical computational structure for emotions identifications. The proposed method follows binary classification, mapping input speech signals in one of the corresponding emotions classes. The concept used at different level in tree is to solve the classification task in easiest way to diminish error propagation. AIBO and USC IEMOCAP databases are used to evaluate the classification method. Using SVM, accuracy archive 72.44% - 89.58%. The results proves the reported hierarchical method is efficient for classifying emotional speech in various databases.

III. PROPOSED ALGORITHM

Aim of the proposed system is to identify emotions from audio files with maximum accuracy. The proposed algorithm consists of following steps:

Step 1: Loading the dataset:
The system uses RAVDESS dataset [15] to identify the emotions. Dataset consists of total 24 professional actors audio files in which 12 are male actor audio and 12 are female actor audio. Audio files includes the various types of emotions as happy, sad, calm, anger, fear, surprise.

Step 2: Plot the basic graph for understanding the audio files:
Understanding of audio files and intensity of audio files we requires to plot the graph(.wav file). In the graph, x axis shows the time(s) and y axis shows the sound amplitude.

Step 3: Emotion labels:
Here are the labels of the emotions category. This creates the dictionary to use when training machine learning model. After creating emotion labels, create the list of emotions which we want to detect. In the proposed system the mainly focused emotions are happy, sad, angry and calm.

Step 4: Extracting Features:
This function will extract the features from audio recordings. The extracted feature are MFCC, Chroma and Mel. MFCC(Mel-Frequency Cepstral Coefficients) is the overall shape of a spectral envelope. Chroma feature relate to the different pitch classes. It captures the harmonic and melodic features.

Step 5: Split the dataset:
The features as input x and labeled emotions as an output y. And then split the dataset using the train_test_split() function. It is function by scikit-learn module. In the model, there are 20% testing data and 80% training data.

Step 6: Initialize MLP Classifier:
A Multi-Layer Perceptron(MLP) is a feedforward artificial neural network that generates a set of outputs from set of inputs. This system uses MLPClassifier for classification of outputs over the number of inputs. MLPClassifier uses a back propopagation for training the network. In the proposed model, 500 initial hidden layers are present.

Step 7: Accuracy Score:

After initializing MLPClassifier and fitting the model, next step is to predict the accuracy score. In the proposed system, the predicted value is stored into the variable called y_pred. Accuracy function checks how many predicted values are matching with the labeled dataset.

IV. PSEUDO CODE

```
Step 1. Start
Step 2. Import required libraries. Librosa, Soundfile, NumPy, Scikit-Learn, Matplotlib.
Step 3. Load the RAVDESS dataset for ML model. Define a function loading audio data().
x as an input feature and y is the labeled emotions as an output.
Step 4. Plot the basic graph for understanding the audio files by importing librosa library.
        Xaxis : Time(s) Y axis : Sound amplitude
Step 5. Extract features from audio files using def audio features() function. Extracting features
from audio files are MFCC, Mel and Chroma.
Step 6. Split the dataset using the train test split() function.
Step 7. Label the classified emotions in the audio files. emotion labels=
Step 8. Initialize MLP ClassifierStep
Step 9. Train the model and fitting the model. model.fit(x train,y train)
Step 10. Assign the predicted values into a new variable y pred. (Detect the emotions)
Step 11. Calculate accuracy score of the prediction using accuracy() function.
Step 12: Stop
```

V. RESULTS

The simulation studies involve deterministic neural network with 500 hidden layers. In proposed speech recognizing algorithm, Multi-Layer Perceptron (MLP) classifier algorithm is implemented in Python with Jupyter Notebook. The system take .wav audio file as input from the RAVDESS dataset [15]. The sample signal waveform for one of the audio file is as shown in Fig.1 and Fig. 2. The input file is given for feature extraction. The extracted features from the given sample audio fileare as shown inFig. 3.

The Extracted features are given to 500 hidden layers of MLP Classifier algorithm. For output calculation Adaptive Activation function is used. Finally the emotion is recognized according to the given audio file. The accuracy of the system is calculated using accuracy-score built-in function from sklearn.matrics module. The experimental resultgives the average accuracy of the system between 59% to 79% as shown in Fig. 4.The averageaccuracy of proposed system is 69.27%. The user is treated according to his emotion/(s) through YouTube videos.The appearance of emotions from speech audio file is shown in Fig. 5.

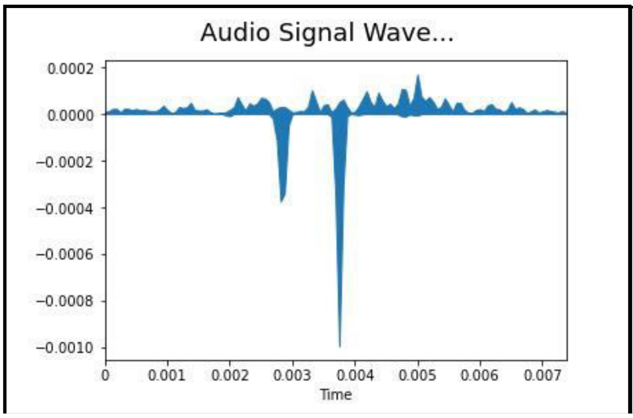


Fig.1. Signal waveform for sample audio file

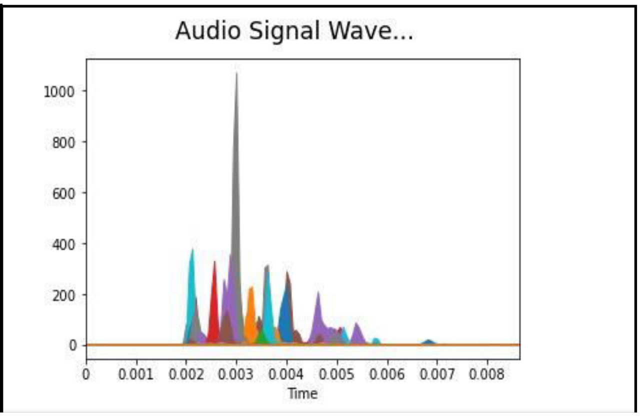
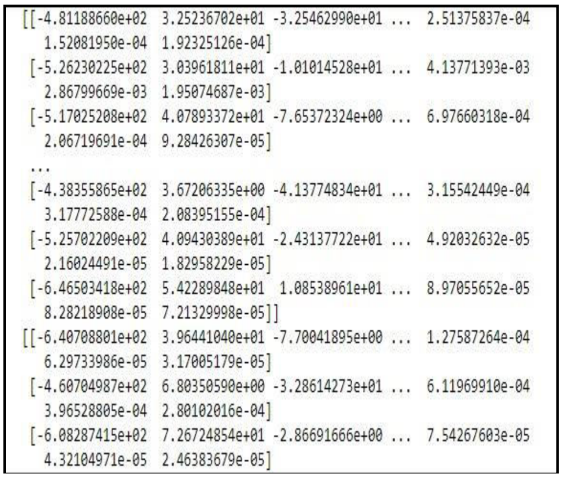
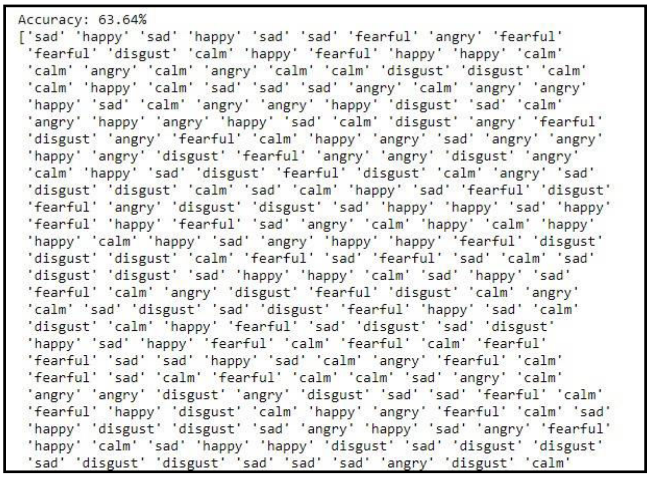


Fig. 2. No of Extracted features from sample audio file



```
[[-4.81188660e+02  3.25236702e+01 -3.25462990e+01 ...  2.51375837e-04
   1.52081950e-04  1.92325126e-04]
 [-5.26230225e+02  3.03961811e+01 -1.01014528e+01 ...  4.13771393e-03
   2.86799669e-03  1.95074687e-03]
 [-5.17025208e+02  4.07893372e+01 -7.65372324e+00 ...  6.97660318e-04
   2.06719691e-04  9.28426307e-05]
 ...
 [-4.38355865e+02  3.67206335e+00 -4.13774834e+01 ...  3.15542449e-04
   3.17772588e-04  2.08395155e-04]
 [-5.25702209e+02  4.09430389e+01 -2.43137722e+01 ...  4.92032632e-05
   2.16024491e-05  1.82958229e-05]
 [-6.46503418e+02  5.42289848e+01  1.08538961e+01 ...  8.97055652e-05
   8.28218908e-05  7.21329998e-05]]
[[-6.40708801e+02  3.96441040e+01 -7.70041895e+00 ...  1.27587264e-04
   6.29733986e-05  3.17005179e-05]
 [-4.60704987e+02  6.80350590e+00 -3.28614273e+01 ...  6.11969910e-04
   3.96528805e-04  2.80102016e-04]
 [-6.08287415e+02  7.26724854e+01 -2.86691666e+00 ...  7.54267603e-05
   4.32104971e-05  2.46383679e-05]
```

Fig. 3. Extracted Features and Emotions



```
Accuracy: 63.64%
['sad' 'happy' 'sad' 'happy' 'sad' 'sad' 'fearful' 'angry' 'fearful'
 'fearful' 'disgust' 'calm' 'happy' 'fearful' 'happy' 'happy' 'calm'
 'calm' 'angry' 'calm' 'angry' 'calm' 'calm' 'disgust' 'disgust' 'calm'
 'calm' 'happy' 'calm' 'sad' 'sad' 'sad' 'angry' 'calm' 'angry' 'angry'
 'happy' 'sad' 'calm' 'angry' 'angry' 'happy' 'disgust' 'sad' 'calm'
 'angry' 'happy' 'angry' 'happy' 'sad' 'calm' 'disgust' 'angry' 'fearful'
 'disgust' 'angry' 'fearful' 'calm' 'happy' 'angry' 'sad' 'angry' 'angry'
 'happy' 'angry' 'disgust' 'fearful' 'angry' 'angry' 'disgust' 'angry'
 'calm' 'happy' 'sad' 'disgust' 'fearful' 'disgust' 'calm' 'angry' 'sad'
 'disgust' 'disgust' 'calm' 'sad' 'calm' 'happy' 'sad' 'fearful' 'disgust'
 'fearful' 'angry' 'disgust' 'disgust' 'sad' 'happy' 'happy' 'sad' 'happy'
 'fearful' 'happy' 'fearful' 'sad' 'angry' 'calm' 'happy' 'calm' 'happy'
 'happy' 'calm' 'happy' 'sad' 'angry' 'happy' 'happy' 'fearful' 'disgust'
 'disgust' 'disgust' 'calm' 'fearful' 'sad' 'fearful' 'sad' 'calm' 'sad'
 'disgust' 'disgust' 'sad' 'happy' 'happy' 'calm' 'sad' 'happy' 'sad'
 'fearful' 'calm' 'angry' 'disgust' 'fearful' 'disgust' 'calm' 'angry'
 'calm' 'sad' 'disgust' 'sad' 'disgust' 'fearful' 'happy' 'sad' 'calm'
 'disgust' 'calm' 'happy' 'fearful' 'sad' 'disgust' 'sad' 'disgust'
 'happy' 'sad' 'happy' 'fearful' 'calm' 'fearful' 'calm' 'fearful'
 'fearful' 'sad' 'sad' 'happy' 'sad' 'calm' 'angry' 'fearful' 'calm'
 'fearful' 'sad' 'calm' 'fearful' 'calm' 'calm' 'sad' 'angry' 'calm'
 'angry' 'angry' 'disgust' 'angry' 'disgust' 'sad' 'sad' 'fearful' 'calm'
 'fearful' 'happy' 'disgust' 'calm' 'happy' 'angry' 'fearful' 'calm' 'sad'
 'happy' 'disgust' 'disgust' 'sad' 'angry' 'happy' 'sad' 'angry' 'fearful'
 'happy' 'calm' 'sad' 'happy' 'happy' 'disgust' 'sad' 'disgust' 'disgust'
 'sad' 'disgust' 'disgust' 'sad' 'sad' 'sad' 'angry' 'disgust' 'calm'
```

Fig 4.Extracted Features and Emotions

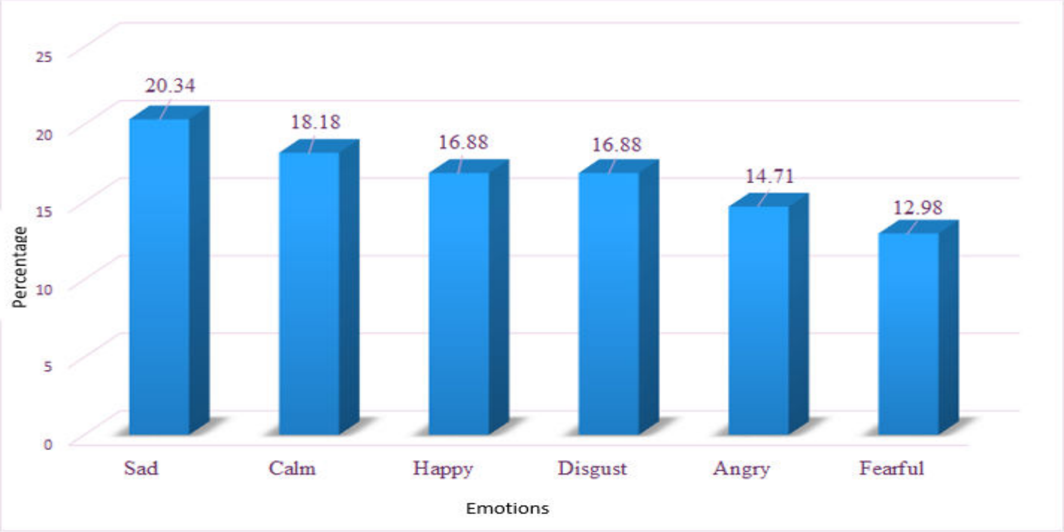


Fig. 5. Appearance of emotions in speech audio file

VI. Conclusion and Future Work

The simulation result has provided a detailed review of the machine learning techniques used for recognizing emotions in speech. Proposed machine learning methods and their layer-wise architectures are briefly elaborated based on the classification of various natural emotion such ashappiness, joy, sadness, neutral, surprise, boredom, disgust, fear and anger. The methods offer easy training of machine learning model as well as efficiency of shared weights.

In future, the speech signals can be analyzed in real time. The emotions can also be detected with other languages. A heuristic approach may be used to further improve the accuracy for machine learning model.

REFERENCES

1. Khalil, Ruhul Amin, Edward Jones, Mohammad Inayatullah Babar, Tariqullah Jan, Mohammad HaseebZafar, and ThamerAlhussain. "Speech emotion recognition using deep learning techniques: A review." *IEEE Access* 7 (2019): 117327-117345.
2. Anagnostopoulos, Christos-Nikolaos, TheodorosIliou, and IoannisGiannoukos. "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011." *Artificial Intelligence Review* 43, no. 2 (2015): 155-177.

3. El Ayadi, Moataz, Mohamed S. Kamel, and FakhriKarray. "Survey on speech emotion recognition: Features, classification schemes, and databases." *Pattern recognition* 44, no. 3 (2011): 572-587.
4. Shi, Peng. "Speech emotion recognition based on deep belief network." In *2018 IEEE 15th International Conference on Networking, Sensing and Control (ICNSC)*, pp. 1-5. IEEE, 2018.
5. Cao, Houwei, RaginiVerma, and AniNenkova. "Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech." *Computer speech & language* 29, no. 1 (2015): 186-202.
6. Nwe, Tin Lay, Say Wei Foo, and Liyanage C. De Silva. "Speech emotion recognition using hidden Markov models." *Speech communication* 41, no. 4 (2003): 603-623.
7. Chen, Lijiang, Xia Mao, YuliXue, and Lee Lung Cheng. "Speech emotion recognition: Features and classification models." *Digital signal processing* 22, no. 6 (2012): 1154-1160.
8. Wu, Siqing, Tiago H. Falk, and Wai-Yip Chan. "Automatic speech emotion recognition using modulation spectral features." *Speech communication* 53, no. 5 (2011): 768-785.
9. Albornoz, Enrique M., Diego H. Milone, and Hugo L. Rufiner. "Spoken emotion recognition using hierarchical classifiers." *Computer Speech & Language* 25, no. 3 (2011): 556-570.
10. Rong, Jia, Gang Li, and Yi-Ping Phoebe Chen. "Acoustic feature selection for automatic emotion recognition from speech." *Information processing & management* 45, no. 3 (2009): 315-328.
11. Wu, Chung-Hsien, and Wei-Bin Liang. "Emotion recognition of affective speechbased on multiple classifiers using acoustic-prosodic information and semantic labels." *IEEE Transactions on Affective Computing* 2, no. 1 (2010): 10-21.
12. Lee, Chul Min, and Shrikanth S. Narayanan. "Toward detecting emotions in spoken dialogs." *IEEE transactions on speech and audio processing* 13, no. 2 (2005): 293-303.
13. Yang, Bin, and Marko Lugger. "Emotion recognition from speech signals using new harmony features." *Signal processing* 90, no. 5 (2010): 1415-1423.
14. Lee, Chi-Chun, Emily Mower, Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. "Emotion recognition using a hierarchical binary decision tree approach." *Speech Communication* 53, no. 9-10 (2011): 1162-1171.
15. Livingstone, Steven R 2018,*RAVDESS Emotional speech audio: Emotional speech dataset* accessed 15 March 2021, < https://www.kaggle.com/uwrfkaggler/ravdess-emotional-speech-audio>.

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING