Speech Emotion Recognition using Multilayer Perceptron Classifier on Ravdess Dataset

Rekha R[1], Tharani R S[2]
{rra.it@psgtech.ac.in[1], 20ma01@psgtech.ac.in[2]}

[1]*Assistant Professor, Department of IT, PSG College of Technology*, [2]*PG Scholar, Department of IT, PSG College of Technology*

Abstract. Emotion is the best way to express one person's thought and action to others. Identifying emotions from one person's speech is the most required technology for today's world. Emotion Recognition can be much helpful to derive various useful insights about the thoughts of a person. Speech emotion recognition (SER) is the process of obtaining emotions like happiness, sadness, neutral and other emotions from one person's speech. In this paper, Speech emotion recognition method is used to gain emotions from RAVDESS dataset. The emotion extraction is done based on speech features like Mel-frequency cepstrum coefficients (MFCC), chroma and mel. The extracted data is then trained with Multilayer perceptron classifier (MLP) and obtained an accuracy of 82.8%..

Keywords: SER, MFCC, Chroma, mel, MLP.

1 Introduction

Emotion is one of amode in which one person expresses his/her ideas and thoughts to others. Speech emotion recognition is the best method for recognising emotions. SER is the method of gaining emotions from one speaker's speech to identify their ideas and thoughts.It mainly identifies emotions like happiness, sadness, neutral, surprise, anger, fear, disgust and many other emotions that can be identified from the speech obtained. To identify emotions from the dataset many speech features like MFCC, Chroma, energy, mel spectrogram, pitch and many more has been used to extract the features. Later classifier models like Support vector machine, Multilayer perceptron classifier, Gaussian mixture model, Hidden Markov model, Naïve bayes classifier, Decision tree, K- nearest neighbourand many other classifier models has been used for recognizing emotions.

2 Literature Survey

Many previous literature papers have used Speech emotion recognition method to obtain emotions from speech. Some of the literature papers are discussed below.

Mansour Sheikhan et al., [2013] proposed a SVM based classifier model on FARSDAT speech corpus dataset and used ANOVA, a technique to analyze data by classifying under various conditions for feature extraction and achieved accuracy of about 80% [1].Wei Gong et

ICCAP 2021, December 07-08, Chennai, India

DOI 10.4108/eai.7-12-2021.2314726

al., [2014] used Buaa emotional speech database for SER and performed feature extraction by using Deep Belief Networks (DBNs) and trained SVM classifier model and achieved accuracy of 86.5% [2].

Mustaqeem and Soonil Kwon [2020] identified different typed of emotion using Gated recurrent units (GRUs)of IEMOCAP and RAVDESS datasets and used Convolutional Long Short-Term Memory (ConvLSTM) for feature extraction. This system obtained an accuracy of 75% [3]. Sanjita. B. R et al., [2020] used MLP classifier over RAVDESS dataset and extracted features like MFCC, chroma and mel features and obtained accuracy of 100% [4].Jerry Joy et al., [2020] proposed MLP classifier and neural network over RAVDESS dataset by using features like MFCC, Contrast, Mel Spectrograph Frequency, Chroma and Tonnetz and obtained accuracy of 70.28% [5]. Misbah Farooq et al., [2020] applied SVM and MLP over EMODB, SAVEE, IEMOCAP and RAVDESS dataset and used Correlation based feature selection (CFS) for feature extraction [6].NimaTaherinejad et al., [2016] used Berlin Emotional database and extracted features like pitch related features, MFCC and applied accuracy of about 70% by applying Naïve bayes classifier [7].

Linhui Sun et al., [2019] applied Decision tree based SVM over Chinese emotion database of the Chinese Academy of Sciences and obtained accuracy of 75.8% [8]. Tin Lay New et al., [2011] extracted InstEner, SyllEner, InstPtch, SyllPtch features from INTERFACE Emotional Speech Synthesis Database and gaines accuracy of 78% by training Hidden Markov model over the data [9]. Ali Meftah et al., [2021] used KSUEmotions corpus dataset and extracted features like zero-crossing rate, short-term energy, MFCC's and delta features. Then they applied SVM and KNN over the data and obtained better accuracy for KNN than SVM [10].

3 Proposed Methodology

In this section, the dataset used, features used for extraction, classifier model and the workflow of obtaining emotions from the dataset is mentioned.

A. Dataset Description

The dataset used in this paper is RAVDESS dataset. RAVDESS dataset is Ryerson Audio-Visual Database of Emotional Speech and Song dataset which consists of 7356 speech audio files in the form of .wav file. It includes 24 actors (12 male, 12 female) and includes emotions like calm, happy, sad, angry, fearful, surprise, and disgust expressions. Each file name is of seven number format which describes modality, vocal channel, emotion, emotional intensity, statement, repetition, actor.A total of 7356 audio files are available in the dataset. The audio files are evaluated by 247 persons for about ten times to verify the correctness of the labels given to the audio files with respect to the emotions. The emotions are labelled from 1 to 8 for each emotion and for actor, odd number is for male and even number is for female.

B. Feature Extraction

Audio files can be either speech or song or video files with sound. For extracting emotions from these audio files certain features has to be extracted from the audio files to perform analysis over the dataset. The features considered in this paper are MFCC, Chroma and mel.

- ***MFCC***

Mel Frequency Cepstral Coefficients(MFCC) is an audio feature which is mainly used for feature extraction purpose. MFCC is obtained from audio signals by breaking the signal into

overlapping frames and then fast fourier transform is applied over the signals. Then sent to filter bank for removing noise and cepstral coefficients are obtained.

- ***Chroma***

Chroma is also a feature mainly used for extracting features from audio or speech files. Chroma is based on pitch classes where the pitch classes include 12 categories. Chroma consists of two factore they are chroma vector and chroma deviation.

- ***Mel***

The Mel scale relates evident repeat, or pitch, of an unadulterated tone to its real assessed recurrence. Individuals are incredibly improved at perceiving little changes in pitch at low frequencies than they are at high frequencies. Solidifying this scale makes our features arrange even more eagerly what individuals listen.

C. MLP Classifier

Multilayer Perceptron Classifier (MLP) is a class of feed forward Artificial Neural Network (ANN) and is shown in figure 1. It is loosely for feed forward ANN and strictly for network of multiple layers of perceptron. The input features are accepted by the input layer and passed over several hidden layers and finally the classification output is obtained at the output layer.

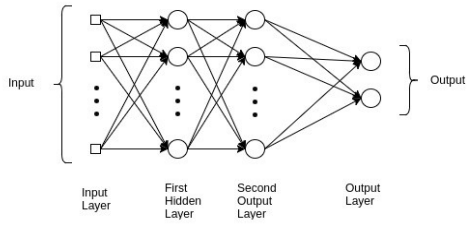

Figure1. Multilayered Perceptron.

D. Workflow

The workflow of proposed research work is shown in figure 2.

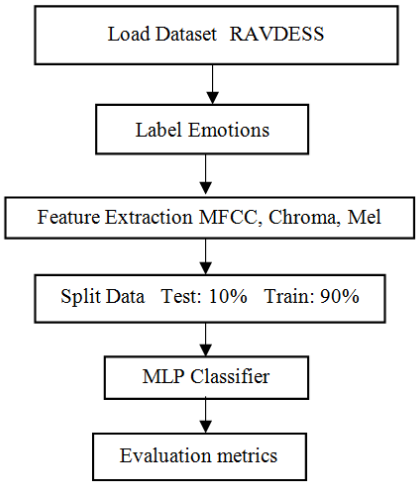

Figure2.Workflow of SER using MLP Classifier

MLP classifier uses supervised learning technique called backpropagation for training. It can be used to extract patterns and detect trends that are too complex for human or computer. It can distinguish data that is not linearly separable. It also has the ability to separate the filtered data using more complex shapes. This model is suitable for data with complex structures and it have an internal neural network for the purpose of performing classification.

4 Implementation

A. Load Dataset

In this paper, initially RAVDESS dataset is downloaded from Kaggle dataset and extracted in the file. This file dataset includes 24 folders with 1440 audio files total. This dataset is loaded for performing analysis. For loading dataset librosa file available in python language is used. Librosa library is mainly used for accessing and analyzing audio-based files.

B. Label Emotions

The RAVDESS dataset includes universal emotions including neutral, calm, happy, sad, angry, fearful, disgust, surprised. These emotions are determined in the dataset based on the file name where it is numbered from 1 to 8. To make speech emotion recognition analysis over dataset emotions play an important role. So, the emotions are labelled as given in the below Table 1.

Label	Emotion
01	Neutral
02	Calm
03	Happy
04	Sad
05	Angry
06	Fearful
07	Disgust
08	surprised

Table 1 Emotion Labels

In RAVDESS dataset only sad, angry andhappy emotions are considered. The waveplot and spectrogram of sad, angry and happy emotions are shown in figure 3, 4 and 5 respectively. The total number of files for the three emotions in the dataset are 192 audio files for happy, 192 audio files for sad, 192 audio files for angry. The total number of features and emotion to be extracted from the dataset is 576.The total count of all emotions is given in Figure.6.

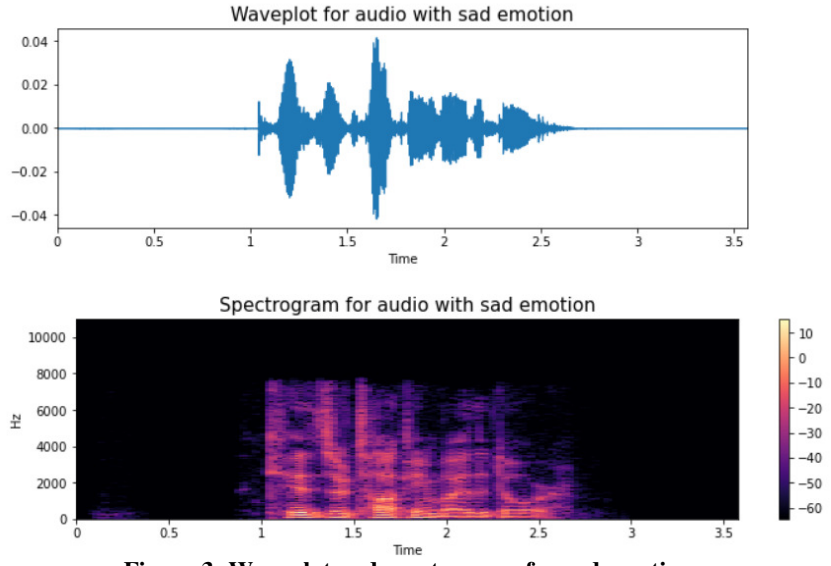

Figure 3: Waveplot and spectrogram for sad emotion

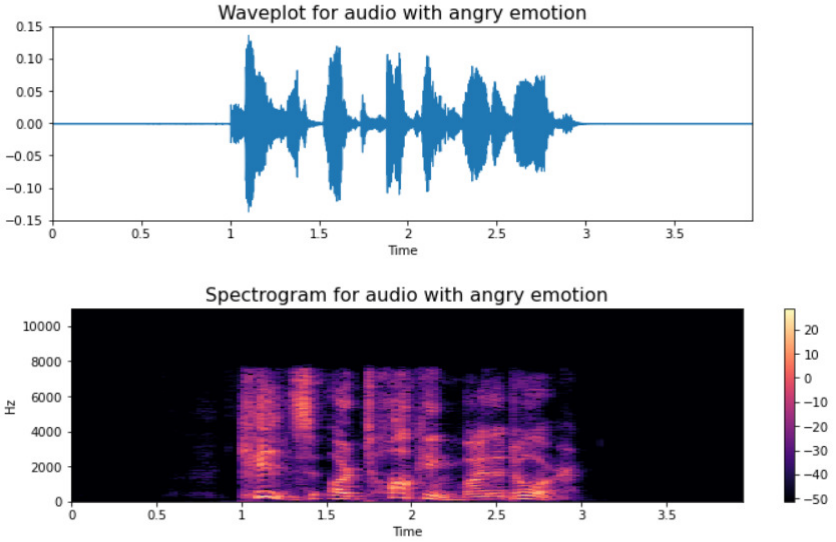

Figure 4: Waveplot and spectrogram for angry emotion

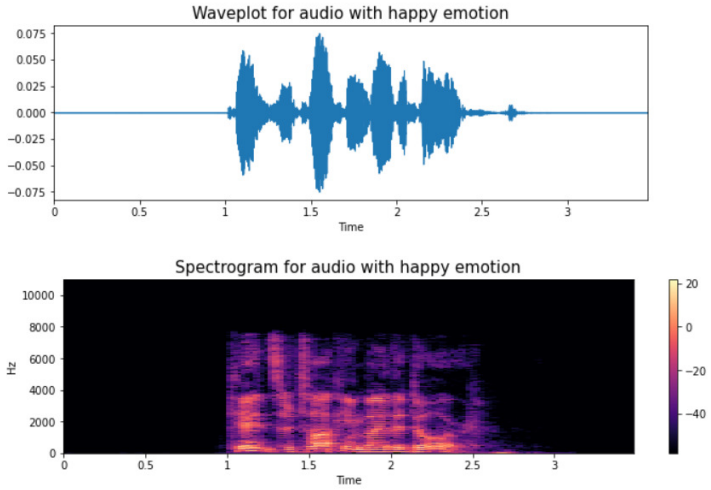

Figure 5: Waveplot and spectrogram for happy emotion

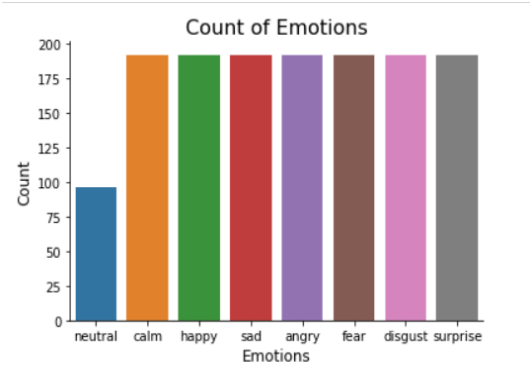

Figure.6. Count of emotions in dataset

C. Feature Extraction using MFCC, Chroma and Mel

In Speech emotion recognition, feature extraction plays a very important role to gain various audio features to perform emotion analysis. In this paper, feature extraction is mainly done based on features like Mel frequency Cepstral Coefficient (MFCC), Chroma and Mel spectrogram. Thesefeatures are extracted by using in-built functions available in librosa and stored in arrays for classification and analysis. In RAVDESS dataset, feature extraction is calculated for 576 audio files according to their emotions. For each file 45 features and 1 emotion label are identified.

D. Classification using MLP Classifier

Multi-Layer Perceptron model is considered to the suitable model for speech emotion recognition. This is because, MLP classifier is best suitable for complex structures datasets when compared to other models like SVM, KNN etc. MLP classifier is an in-built classifier model available as default in scikit learn library which contains many in-built classifier models. The output dataset from the feature extraction module is given as input to MLP classifier model. For performing classification, the dataset is initially splitted into test and train

dataset. 10% of original dataset is used for test and remaining of dataset is used for training. The MLP classifier model is trained over both test and train dataset and their accuracy is compared with each other. Since MLP classier is mainly known for hidden layers, in this paper 100 hidden layers were included with a maximum iteration of 400 iterations. These hiddenlayers are high in-order to increase the accuracy and performance of the model.This MLP classifier classifies different emotions that includes only happy, sad, angry emotions in the dataset.

E. Evaluation metrics

Evaluation metrics says about the different ways of testing the performance of the machine learning model for the given dataset. Selection of appropriate evaluation metric is very important for proper understanding of a model.Accuracy, Precision, Recall and F1 score are the metrics used for the evaluation of the speech emotion recognition model constructed.

	Test	**Train**
Accuracy	87.9%	98.8

Table 2 Accuracy for Test and Train dataset

	Happy	**Angry**	**Sad**
Precision	95%	100%	65%
Recall	76%	95%	100%
F1 score	84%	98%	79%

Table 3 Precision, Recall, F1 score for Test dataset

	Happy	**Angry**	**Sad**
Precision	100%	100%	97%
Recall	99%	98%	100%
F1 score	99%	99%	98%

Table 4 Precision, Recall, F1 score for Train dataset

These evaluation metrics are available as in-built function classification report and accuracy_score function which is imported from scikit learn library. These evaluation metrics are calculated for both train and test dataset and the values are compared for angry, happy and sad emotions. The evaluation metrics of test and train dataset are recorded as given in Tables 2, 3 and 4.

5 Result Analysis

In this paper, emotions considered are happy, sad and angry. Based on these three emotions result analysis is performed using evaluation metrics. As a result, the RAVDESS

dataset has obtained accuracy of about 87.9% for test dataset and 98.8% for train dataset. Other evaluation metrics like precision, recall, F1 score has been considered for three emotions separately for both test and train dataset. The accuracy of the dataset is compared in Figure 3.The evaluation metrics for happy, angry and sad emotions are calculated for both test and train data as shown in figure 7. The loss obtained from 400 iterations in MLP classifier is plotted in Figure 8.

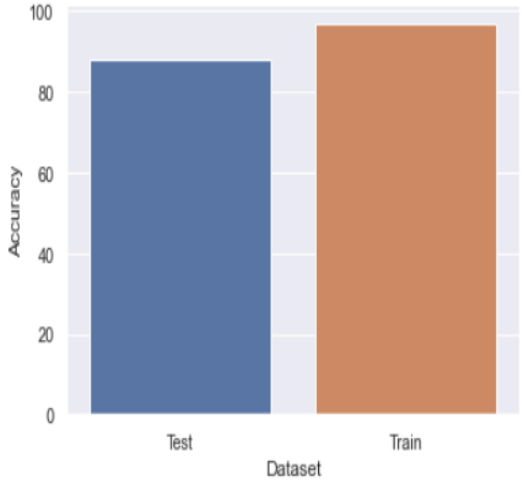

Figure.7. Accuracy of Test and Train data

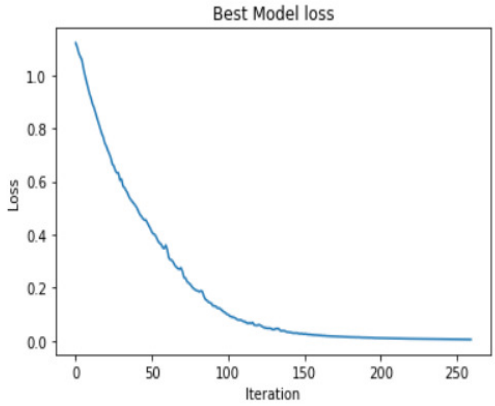

Figure 8. MLP Model loss iteration

6 Conclusion

Among the evaluation metrics calculated for three emotions, in test dataset, Angry has the highest precision of 100%, sad has the highest Recall of 100%and angry has the highest F1 score value of 98%. In train dataset, happy and angry has the same highest precisionof 100%,

sad has the highest recall of 100% and happy and angry has the same highest F1 score of 99%. On comparing the evaluation metrics in test dataset Angry emotion has the highest value and in train dataset Happy emotion has the highest value. This MLP model obtained a high accuracy of 98.8% for train dataset and 87.9% of test dataset. This accuracy may vary according to iterations in MLP classifier.

References

[1] Sheikhan, M., Bejani, M., &Gharavian, D. (2013). Modular neural-SVM scheme for speech emotion recognition using ANOVA feature selection method. Neural Computing and Applications, 23(1), 215-227.

[2] Huang, C., Gong, W., Fu, W., & Feng, D. (2014). A research of speech emotion recognition based on deep belief network and SVM. Mathematical Problems in Engineering, 2014.

[3] Kwon, S. (2020). CLSTM: Deep feature-based speech emotion recognition using the hierarchical ConvLSTM network. Mathematics, 8(12), 2133.

[4] Palo, H. K., Chandra, M., & Mohanty, M. N. (2017). Emotion recognition using MLP and GMM for Oriya language. International Journal of Computational Vision and Robotics, 7(4), 426-442.

[5] Palo, H. K., Mohanty, M. N., & Chandra, M. (2015). Use of different features for emotion recognition using MLP network. In Computational Vision and Robotics (pp. 7-15). Springer, New Delhi.

[6] Farooq, M., Hussain, F., Baloch, N. K., Raja, F. R., Yu, H., &Zikria, Y. B. (2020). Impact of feature selection algorithm on speech emotion recognition using deep convolutional neural network. Sensors, 20(21), 6008.

[7] Urbano Romeu, Á. (2016). Emotion recognition based on the speech, using a Naive Bayes classifier (Bachelor's thesis, UniversitatPolitècnica de Catalunya).

[8] Sun, L., Zou, B., Fu, S., Chen, J., & Wang, F. (2019). Speech emotion recognition based on DNN-decision tree SVM model. Speech Communication, 115, 29-37.

[9] Nwe, T. L., Foo, S. W., & De Silva, L. C. (2003). Speech emotion recognition using hidden Markov models. Speech communication, 41(4), 603-623.

[10] Meftah, A., Qamhan, M., Alotaibi, Y. A., &Zakariah, M. (2020). Arabic Speech Emotion Recognition Using KNN and KSUEmotions Corpus. International Journal of Simulation--Systems, Science & Technology, 21(2).