International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 08 Issue: 04 | Apr 2021 www.irjet.net p-ISSN: 2395-0072

Speech Emotion Recognition using Acoustic Features

Amrutha K[1], Sunanda Panigrahi[2], Rohit M[3], Rama S[4]

[1,2,3]Student, Dept. of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, No.1, Jawaharlal Nehru Road, Vadapalani, Tamil Nadu, India
[4]Assistant Professor, Dept. of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, No.1, Jawaharlal Nehru Road, Vadapalani, Tamil Nadu, India

---

Abstract - *Human-computer interaction is a complex combination of both human emotions and intentions. Emotion recognition from speech signals is a challenging task due to the complexity of human emotions. There is a rise in the use of voice assistants these days, and the significance of using audio data to recognize emotions plays a very important role. In this paper,we propose the recognition of speech emotions based on the acoustic features of the person's voice. The features include Mel Frequency Cepstral Coefficients (MFCC), Chroma Vector and Zero Crossing Rate. In this work, we test the benefit of sound augmentation techniques in speech emotion recognition. We have implemented algorithms like convolutional neural networks (CNN) and Multi-Layer Perceptron (MLP) are used to classify and predict the emotions based on the relevant features extracted from the speech signals. The model is tested on two datasets English datasets, RAVDESS and Surrey Audio-Visual Expressed Emotion (SAVEE). The CNN model has shown 86% accuracy and the MLP model has shown 79% accuracy.*

Key Words: SER, MFCC, Chroma vector, Data Augmentation, CNN, MLP, Zero crossing Rate

1. INTRODUCTION

Emotions play a central role in human-computer interaction, especially when we make use of voice assistants in our daily lives. Proper speech analysis can predict the emotions of an individual reliably. Currently voice assistants are dominating the market and are used even for minuscule tasks. According to current reports [1], half of the global internet users use digital voice assistants with countries like India, Mexico and UAE leading the way. Digital voice assistants have reported a 96% overall satisfaction rate globally.

It was predicted that by the end of 2020, 50% of all searches would be voice activated [2]. Voice assistance for navigation is also an up and coming field and emotion identification can be applied here to monitor the riders' patience rate and can be used to prevent accidents [3]. Identification of emotions with the help of speech analysis can bring great changes to the customer satisfaction [4], companies are currently bringing technologies which can detect the emotional tendencies of an user at different steps of their product usage journey (frustration during setting up the procedure, impatience during delay of installation can be identified easily).

By the current trends, the world is steadily moving towards a speech based help system. In future consumer support,e-commerce systems, daily task scheduling, and everything related to modern living will be shifted towards an automatic procedure. Analysing and detecting emotion can help companies improve their services and ensures higher customer satisfaction

In this paper, we consider 8 emotion classes namely : Happy, Sad, Neutral, Angry, Fear, Calm, Surprised and Disgust and try to detect the above emotions from the audio files. The Speech Emotion Recognition system is trained using a combination of two databases RAVDESS and SAVEE. Once the audio files have been pre-processed, the important acoustic attributes such as MFCC, Chroma and Zero Crossing Rate are extracted. Sound Augmentation is done before feature extraction to enhance the audio features. Classification of emotions is performed using two algorithms, Convolution Neural Networks (CNN) and Multi-Layer Perceptron (MLP).

This paper is organized as follows, Section 2 describes the literature survey, Section 3 elaborates on the proposed methodology, the datasets used and the features extracted. Section 4 describes the model architecture, and classification algorithms implemented. Section 5 is the results and discussion and Section 6 is the conclusion and future scope.

2. LITERATURE SURVEY

Researchers have developed many different methods of SER. Among most research, the most common dataset used is RAVDESS and the standard feature used is MFCC. [5] proposes a multi-modal method for classifying emotions on the RAVDESS dataset. They use Modulation Spectral (MS) and 10 Mel-frequency cepstrum coefficients (MFCC) features on different classifiers such as CNN, SVM, RF, and decision tree of which CNN had the best performance with accuracy of 78%. [6] used KKN to classify seven emotions from the Berlin (EMO-DB) and Spanish (SES) corpora by implementing feature subset selection methods. They used Sequential Forward Selection and Sequential Floating Forward Selection to subset the most informative features. The features used in

this paper are MFCCs, and statistical features like mean, variance, range, skewness, kurtosis.

[7] focuses on the conduction of speech emotion recognition experiment using real voice messages. They created a custom dataset with real WhatsApp voice messages of participants. From this dataset, the obtained features such as MFCCs, chroma, Time-domain cues and frequency domain cues. The model is implemented on three classifiers namely, SVM, KNN, and MLP. The paper [8] focuses on increasing the performance of classifiers using dimension reduction algorithms such as Principal Component Analysis, Recursive Feature Elimination. The proposed model detects five emotions using SVM, RF, and Gradient Boosting on the RAVDESS dataset with an accuracy of 60%, 62% and 63% respectively.

A cross-corpus multilingual SER was proposed by [9]. They have used four datasets, namely the SAVEE, URDU, EMO-DB, and EMOVO dataset. The model trains on one dataset and tests on another dataset. This is done to generalize the model to accommodate multilingual environments. The classifiers used in this model are SVM and Random forest. The experiments showed an increase of accuracy using this method. [10] classifies seven emotions from the Berlin and Spanish datasets. Three ML algorithms are used to classify the seven emotions after extracting the acoustic features: Recurrent Neural networks, SVM and MLR. The SER got highest accuracy of 94% on Spanish data using RNN .University of California Interactive Emotion Capture Dataset(IEMOCAP) was used to make a model to classify 4 emotions [11]. The Acoustic and Lexical features were extracted and used for classification using algorithms such as Hidden Markov Models , Gaussian Mixture Models (GMM), artificial neural networks (ANN), K-nearest neighbors (KNN), and support vector machines. An accuracy of 69.2% was observed.

[12] uses Convolutional Neural Network wherein the audio samples are converted into 2D array, which are then modified into short time Fourier, transform (STFD) which are then processed using Recurrent Neural Network. [13] proposed a two-stage feature selection method. In the first stage of selecting the features, the selected features are then fused together for speech emotion recognition. In second stage feature selection, optimal feature subset selection techniques are used to eliminate the curse of dimensionality. Linear Discriminant Analysis (LDA), Regularized Discriminant Analysis (RDA), Support Vector Machine (SVM) and KNearest Neighbor (KNN) classification have been used. Nearly 70% accuracy had been achieved.

3. PROPOSED METHODOLOGY

3.1 Datasets Used

In our work, we have used two open-sourced english datasets, RAVDESS and SAVEE. Our proposed system predicts 8 emotions, namely Anger, Happy, Sad, Fear, Calm, Neutral, Surprise and Disgust. We used a fusion of both datasets to train our model. The Surrey Audio-Visual Expressed Emotion (SAVEE) database has been developed for the growth of automatic emotion recognition systems[14]. The database consists of recordings from 4 male actors with 7 different emotions (neutral, anger, disgust, fear, happiness, sadness and surprise).

Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [15] is a multi modal database containing both audio and video.The dataset consists of 8 emotions (neutral, happy, sad, calm, angry, fearful, disgust and surprise) and 12 male actors and 12 female actors.

The total of 1920 audio files were used in training and testing the speech emotion recognition model.

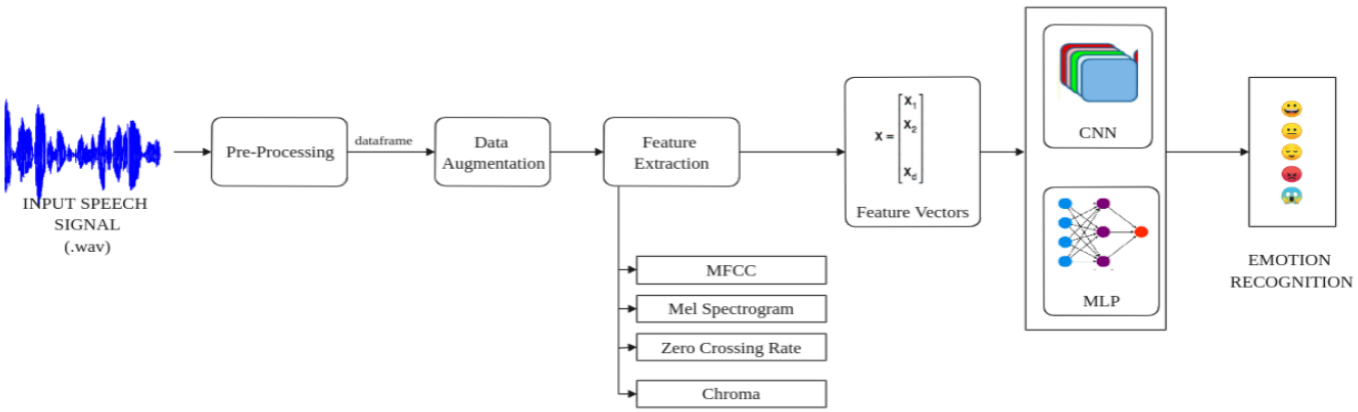

Fig-1: Proposed Architecture

3.2 Feature Extraction

A sound consists of many different kinds of features like frequency, pitch, tone, amplitude and feature extraction is the most important step to analyse it. The wav files are first converted into an array containing the samples of amplitude and the sample rate, which is the number of samples recorded per second. This is then used to extract acoustic features. In our paper, we have used Mel Frequency Cepstral coefficient (MFCC), Chroma Vector, and Zero-Crossing Rate.

3.2.1 Mel Frequency Cepstral Coefficient (Mfcc)

Mel Frequency Cepstral Coefficients help in efficiently extracting the right emotional state of the speech sample. It is the most used feature used in speech applications as it models the human perception of the frequency of speech well [16]. The Mel-scale is used to match the frequency perceived by human ears with the actual frequency. The MFCC is calculated by splitting the audio into multiple frames. Then Fourier transform and power spectrum are calculated for each frame and related to the Mel-Scale. The discrete cosine transform (DCT) is calculated on the Mel log energies and the coefficients are estimated[10]. In our work, we extracted the first 13 MFCCs.

3.2.2 Chroma Vector

The chroma vector relates the twelve different pitch classes. The main characteristic of chroma features are its robustness to the variations in timber and also closely correlate to the musical aspect of harmony. In the chroma circle, the two octave related pitches will have the same angle. This important relation is not captured by a Mel Scale or even a linear pitch scale.

3.2.3 Zero-crossing Rate

Zero Crossing Rate is the rate of sign change of the signal in a frame. It is calculated by dividing the number of times the signal changes from positive to negative with the length of the frame. ZCR can also be used to measure the noisiness of a signal.

4. IMPLEMENTATION

In this paper, we aim to make a novel methodology by experimenting the effect of sound augmentation in Speech Emotion Recognition. Before the extraction of features, we perform Data Augmentation on the audio files. This is done in order to easily distinguish between each of the 8 emotions.

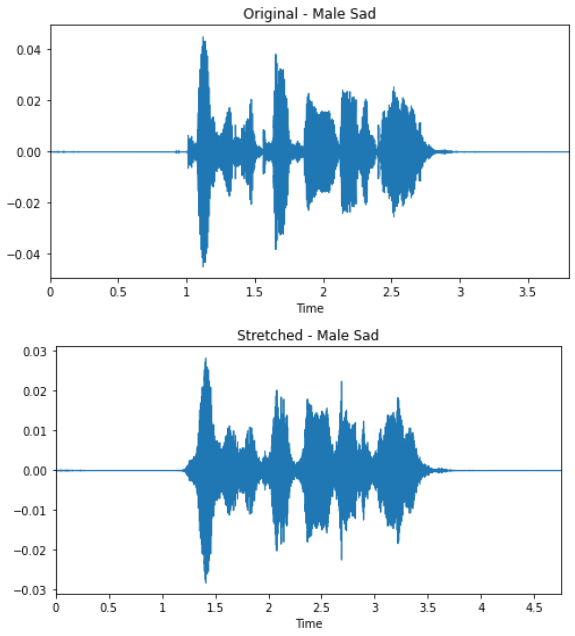

Fig -2 : Spectrograms before and after stretching audio

In our work, we perform some sound augmentation techniques. Stretching is done by changing the duration of the audio without affecting the pitch of the sound. The audio files were stretched by 80% to clearly capture the acoustic features. We also increased the speed of the audio files by a factor of 1.4, and along with the speed, the pitch also increases. Once we have the two augmented audio files, we proceed to extracting the features. The extracted features are then put into feature vectors which will be used for classification. According to our study of various models implemented in SER systems, CNN and MLP have shown to be some of the best performing models [5] [17]. Hence we have used two classifiers, Convolution Neural Network and Multilayer Perceptron to test the effect of sound augmentation in predicting the emotions and compared their performances.

4.1 Convolution Neural Network

Convolution Neural Networks (CNN) are a variant of the standard neural network which is used for recognition of patterns and data classification. It is used for analysing data of the form of a multidimensional array [18].A typical CNN architecture consists of a convolution layer, Pooling layer and fully connected layers.

In our work, we created a CNN model with three convolution layers. Convolution layer is used to extract the features using an element wise multiplication between input data and a filter or a kernel. Here, we have used a kernel of size 10. The activation function used is ReLU. Then we select the best extracted features into an 8*8 matrix using the Max Pooling layer. In order to reduce overfitting, we have also implemented bias and kernel regularizers. The kernel regularizer is used to reduce the

weights and the bias regularizer is used to reduce the bias. In both these regularizers, we opted for L2 regularization. A dropout of probability of 0.4 was added inorder to disperse the parameter values among the nodes. The final dense layer is used to give the output by flattening the multidimensional vector into a single vector of 8 emotions. The total trained parameters were 609,992. The model shows an accuracy of 86.35% in predicting the emotions. The Fig. 3. shows the CNN model's accuracy and loss. The training accuracy of the model reaches close to 100% whereas the testing accuracy is around 86%. The loss function used here is categorical cross entropy. A model's loss is the difference between the predicted value by the model and the true value.

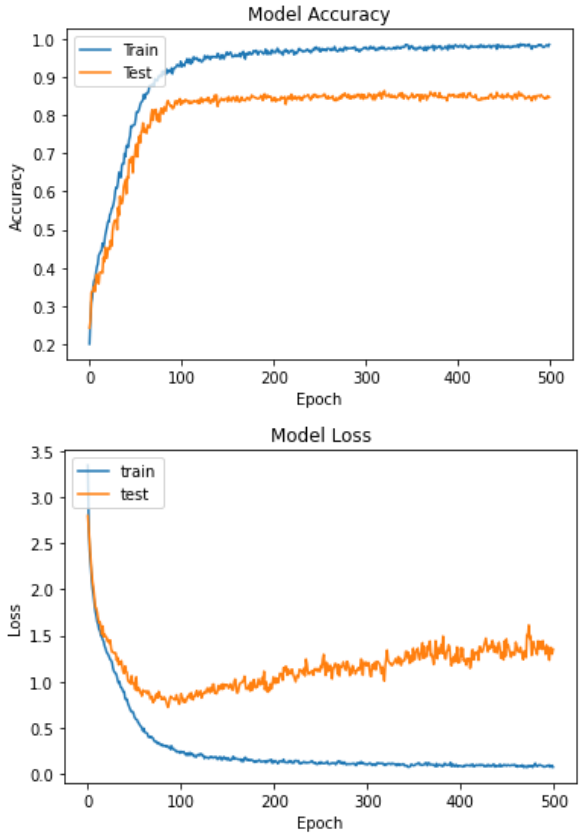

Fig -3: Accuracy and Loss of CNN Model

4.2 Multi-layer Perceptron

A multi-layer perceptron is a feed forward artificial neural network a layer in MLP is made up of one or more perceptrons. In the input layer of a MLP classifier, the input data is used for calculating the weights and passed to the activation layer. This is then used as an input for the next hidden layer. The ith layer neurons are the input to the i+1th layer. The activation function used in this work is rectified linear units (ReLU). This function is repeated till the output layer.

The same dataset was used to train the MLP Classifier and the accuracy reported was 79.34%. After experimenting with different hyper parameter values, the most optimal one was selected and had a batch size of 256 and the hidden layer of MLP classifier was 250. The maximum iteration was set to 600 with an alpha value of 0.01. The MLP model was trained on the same dataset with an adaptive learning rate and ReLU activation function. The observed accuracy was 79.34%

5. RESULTS AND DISCUSSION

The prediction of the emotions on the test data showed an accuracy of 86% for CNN model and 79% for MLP model. The performance of the model is higher than the existing models [5],[16],[19]. This increase in performance can be attributed to data augmentation techniques performed on the audio files before processing it. Fig. 4 and Fig 5 are the confusion matrices of CNN and MLP Model. It gives us a comparison between the correctly classified emotions and the incorrect ones.

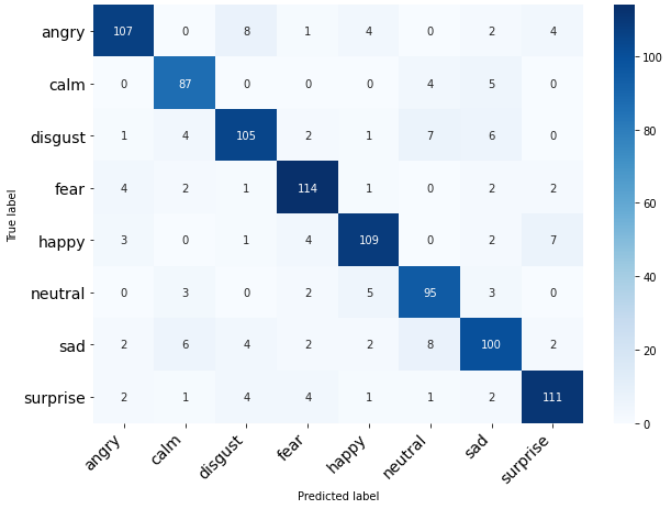

Fig -4: Confusion matrix for CNN Model

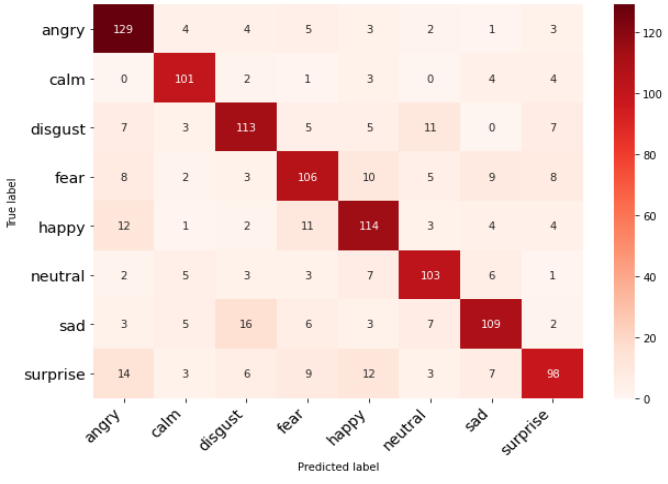

Fig -5: Confusion Matrix for MLP Model

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed a method to recognise emotions in speech using Convolution neural networks and Multilayer Perceptron. The speech includes neutral, calm, happy, angry, fear, sad, surprised and disgust. The audio files first undergo sound augmentation before feature extraction. The relevant features are then used for detecting the emotion. Based on our experiments, the CNN model is better performing compared to the MLP model. Based on the experiments conducted, we can conclude that sound augmentation techniques have a positive impact on the efficacy of the speech emotion recognition system. Speech Emotion recognition has many practical applications and can definitely help to improve human-computer interaction. In the future, we would like to work on multi-lingual datasets and real-time audio detection. The possibilities of the application of speech emotion recognition are endless.

REFERENCES

[1] "Reshape to Relevance." Accessed: Mar. 14, 2021. [Online].Available: https://www.accenture.com/_acnmedia/PDF-93/Accenture-Digital-Consumer-2019-Reshape-To-Relevance.pdf#zoom=50.

[2] Muthukumaran, A.P.K. and Vani, H., 2020. Optimizing the usage of voice assistants for shopping. Indian Journal of Science and Technology, 13(43), pp.4407-4416.

[3] C. D. Katsis, G. Rigas, Y. Goletsis, and D. I. Fotiadis, "Emotion Recognition in Car Industry," in Emotion Recognition, John Wiley & Sons, Ltd, 2015, pp. 515–544.

[4] Ren, Fuji, and Changqin Quan. "Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: an application of affective computing." Information Technology and Management 13, no. 4 (2012): 321-332.

[5] Christy, A., S. Vaithyasubramanian, A. Jesudoss, and MD Anto Praveena. "Multimodal speech emotion recognition and classification using convolutional neural network techniques." International Journal of Speech Technology 23 (2020): 381-388.

[6] S. Kuchibhotla and M. S. R. Niranjan, "Emotional Classification of Acoustic Information With Optimal Feature Subset Selection Methods," Int. J. Eng. Technol., vol. 7, pp. 39–43, May 2018, doi: 10.14419/ijet.v7i2.32.13521.

[7] Gómez-Zaragozá, Lucía, Javier Marín-Morales, Elena Parra, Jaime Guixeres, and Mariano Alcañiz. "Speech Emotion Recognition from Social Media Voice Messages Recorded in the Wild." In International Conference on Human-Computer Interaction, pp. 330-336. Springer, Cham, 2020.

[8] Biswas, Aditi, Sovon Chakraborty, Abu Nuraiya Mahfuza Yesmin Rifat, Nadia Farhin Chowdhury, and Jia Uddin. "Comparative Analysis of Dimension Reduction Techniques Over Classification Algorithms for Speech Emotion Recognition." In International Conference for Emerging Technologies in Computing, pp. 170-184. Springer, Cham, 2020.

[9] Zehra, Wisha, Abdul Rehman Javed, Zunera Jalil, Habib Ullah Khan, and Thippa Reddy Gadekallu. "Cross corpus multi-lingual speech emotion recognition using ensemble learning." Complex & Intelligent Systems (2021): 1-10.

[10] Kerkeni, Leila, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub, and Catherine Cleder. "Automatic speech emotion recognition using machine learning." In Social media and machine learning. IntechOpen, 2019.

[11] Jin, Qin, Chengxin Li, Shizhe Chen, and Huimin Wu. "Speech emotion recognition with acoustic and lexical features." In 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 4749-4753. IEEE, 2015.

[12] Lim, Wootaek, Daeyoung Jang, and Taejin Lee. "Speech emotion recognition using convolutional and recurrent neural networks." In 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA), pp. 1-4. IEEE, 2016.

[13] Kuchibhotla, Swarna, Hima Deepthi Vankayalapati, and Koteswara Rao Anne. "An optimal two stage feature selection for speech emotion recognition using acoustic features." International journal of speech technology 19, no. 4 (2016): 657-667.

[14] Sinith, M. S., E. Aswathi, T. M. Deepa, C. P. Shameema, and Shiny Rajan. "Emotion recognition from audio signals using Support Vector Machine." In 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), pp. 139-144. IEEE, 2015.

[15] Livingstone, Steven R., and Frank A. Russo. "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English." PloS one 13, no. 5 (2018): e0196391.

[16] L. Rabiner and B. Juang, "Fundamentals of Speech Recognition," Prentice Hall Englewood Cliffs N. J., 1993.

[17] Vege, Hari Kiran, and Swarna Kuchibhotla. "MLP Model for Emotion Recognition using Acoustic Features." International Journal 8, no. 5 (2020).

[18] LeCun, Yann, Léon Bottou, Yoshua Bengio, and Patrick Haffner. "Gradient-based learning applied to document recognition." Proceedings of the IEEE 86, no. 11 (1998): 2278-2324.

[19] Sardar, Abdullah Al Mamun, Md Sanzidul Islam, and Touhid Bhuiyan. "A Review on Automatic Speech Emotion Recognition with an Experiment Using Multilayer Perceptron Classifier." In Soft Computing Techniques and Applications, pp. 381-388. Springer, Singapore, 2021.