Comparative Study of Classification Models for Emotion Detection from Speech

[1]Shibraj Basak, [2]Prolay Ghosh

[1]Research Scholar, [2]Assistant Professor
[1]Department of Computer Application,
[2]Department of Information Technology
JIS College of Engineering, Kalyani, India

***Abstract*— The detection of emotions from speech is the aim of this paper. Speech consists of anger, joy and fear have very high and wide range in pitch, whereas Speech consists of sad and tired emotion have very low pitch. Speech Emotion detection technology can recognize human emotions to help machines better for understanding intentions of a user to improve the human-computer interaction. Classification models named Convolutional Neural Network (CNN), Support Vector Machine (SVM), Multilayer Perceptron (MLP) based on mainly Mel Frequency Cepstral Coefficient (MFCC) feature to detect emotion have been presented here. The models have been trained to distinguish eight different emotions such as calm, neutral, angry, sad, happy, disgust, fear, surprise. The proposed work shows that CNN works best on RAVDESS dataset rather than MLP, SVM and records an accuracy of 63.88%.**

***Index Terms*— Emotion Detection, SVM, CNN, MLP, RAVDESS, Machine Learning.**

I. INTRODUCTION

Emotion detection is the process of identifying human emotions using various processes. Different methods have been introduced to detect emotion. Siri, Cortana are very intelligent assistance. Emotion detection can help to improve computer and human interaction by accepting emotions.

Identifying human emotions Accurately, is a difficult task because of their complexness. Some emotions can have different expressions, and some emotions can have similar expressions. Emotions depends on character, culture, gender, situation and locality of a person. Sometimes, it can be hard for a real person to detect emotion from speech of someone. In the case of emotion detection by computers, it is even hard to detect true emotion of a user because of the emotion's complexity. There are different types of systems that use information like audio, text, image, video to detect human emotions, but sometimes not all information is available to use. For example, call centers can use speech recognition to detect emotions.

Detecting emotions from speech is a very tough work. A person's speaking style, volume, intonation, speed, words, etc., greatly affect the detection of emotion in speech. Also, character, culture, gender, situation and locality, etc. are the other factors that can affect the detection of human emotion from speech. Fortunately, several methodologies have been developed to detect human emotions from speech. These methodologies are categorized by their advantages that make them superior to others. In this paper, different approaches have been discussed and compared to find best algorithm among them.

Remaining paper is sectioned as follows: Section II. states the related work. The methodology part includes dataset, feature extraction, algorithms, are introduced in Section III. Section IV shows System design. The results are shown in Section V and conclusions in Section VI.

II. RELATED WORK

The research of emotion recognition from speech has become incredibly popular. Many works have done to advance this field. A work related to this, proposed by Aseef Iqbal and Kakon Barua [1] used Gradient Boosting, Support Vector Machine (SVM) and K-Nearest Neighbor (KNN) classification models to extract real-time emotion from live recorded audio by analyzing tonal properties. RAVDESS (male), RAVDESS (female), RAVDESS (combined) and SAVEE these datasets are used to train this model. This system looks for the four basic emotion classes – anger, sadness, neutral and happiness. In this test, Male testing performance is superior than female in all the classifiers. Compared to other classifiers, gradient boosting has higher classification accuracy across all emotion classes for both males and females. Another related work proposed by Shubham Singh Chaudhary, Sachin Garg [2] used Multilayer Perceptron (MLP), Long short-term memory (LSTM) and Convolutional Neural Network (CNN) classification models to extract emotion from the audio data. Among these classification models CNN model performs better with Kaggle database. Convolutional Neural Network (CNN) model obtained an accuracy of 60%, Long short-term memory (LSTM) model obtained an accuracy of 15% and Multilayer Perceptron (MLP) model obtained an accuracy of 25%.

III. METHODOLOGY

Dataset

This proposed work uses RAVDESS Dataset. This dataset contains 1440 files, recorded by 24 actors, uttering two lexically equivalent sentences with a neutral North American accent. This includes different classes of emotion sad, disgust, surprise, calm, happy, fearful, and angry. Each expression is produced at two levels of emotional intensity (normal, strong), with an additional neutral expression [8].

Data Visualization

Visualization of data provides a greater understanding of the problem and a potential type of solution. Techniques for visualizing the data include classes distribution, instance counts inside each class, the distribution of the dataset, the connection between the characteristics, and dataset clustering. Data visualization functions are available in the Python and R languages.

Data Preparation

It's time to get the data ready for processing once data analyzations are done using different visualizations. The procedures involve in data preparation include correcting difficulties with quality, standardization, and normalization. The data are initially checked for problems including missing values, same type of data, outliers, erroneous data. The dataset did not include any incorrect, identical, or incomplete data.

Data Augmentation

Fresh synthetic data samples can be generated by adding minor changes to our training set, this technique known as data augmentation. Noise input, time altering, pitch and speed changes, and pitch and time switching can be used to provide syntactic data for audio. Making the model resistant to these changes will increase its generalizability. The labels from the initial training are preserved when adding the changes for this to work.

Feature Extraction

Feature extraction is a critical step in researching and identifying connections between numerous things. Since it is already known that the provided audio data cannot be directly processed by the models, they must be converted into a format that can be easily understood, and feature extraction is done to do this. The three axes of the audio signal are time, amplitude, and frequency which represent its three dimensions. In this project, only 5 features are extracted and they are Chroma_stft, Zero Crossing Rate, MelSpectogram, MFCC and RMS (root mean square) value and to train our model.

Modelling

Convolutional Neural Network (CNN), Support Vector Machine (SVM) and Multilayer Perceptron (MLP) classifiers are introduced here to train and test the model. The dataset is split into training and testing data in 3:1 ratio.

IV. SYSTEM DESIGN FLOWCHART

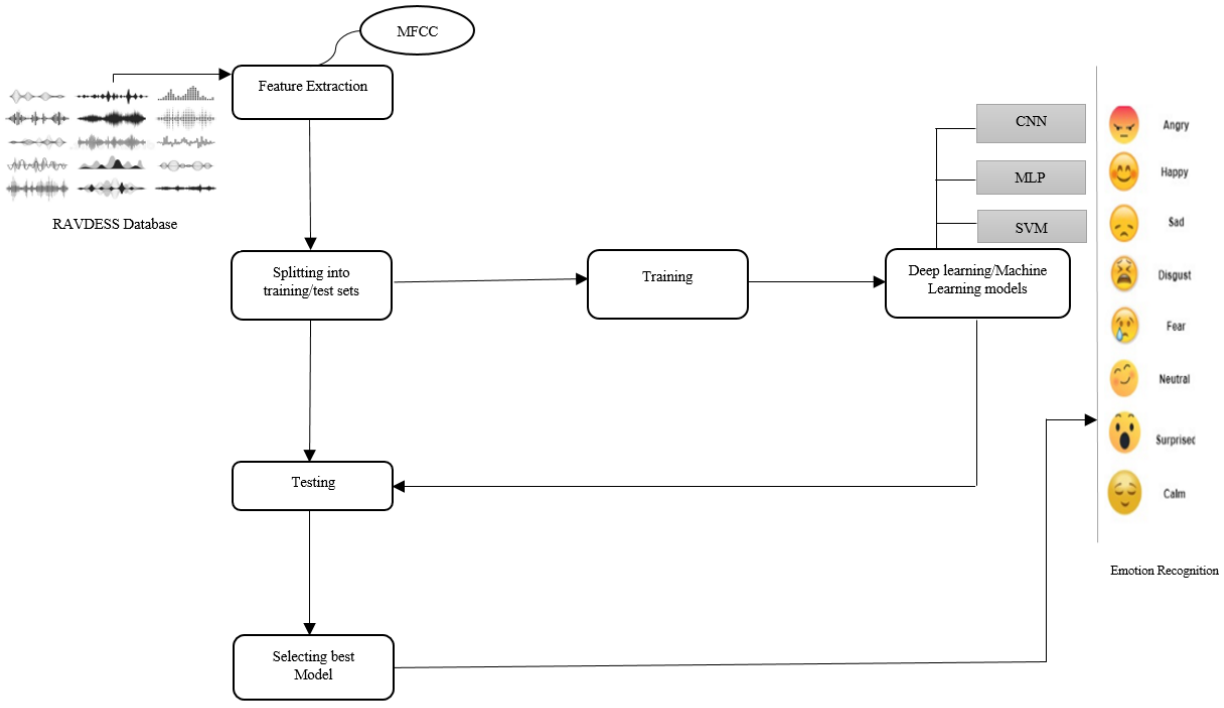

Fig.1: System Design Flowchart

V. RESULT ANALYSIS

In this paper, CNN, SVM and MLP Classification models have been tested to find the best model among them, for RAVDESS Dataset. Table 1 shows accuracy on each of the emotion classes obtained by these three Classification models. Table 2 shows the average accuracy of the used classification models. By comparing them, it has been found that CNN works better with RAVDESS Dataset than SVM and MLP Classification models. Figure 2 shows the Confusion Matrix for CNN model. Figure 3 shows the CNN Model Loss and Accuracy plot against epochs for Training and Testing Data.

Table 1 Accuracy on each of the emotion classes

CLASS	SVM	MLP	CNN
ANGRY	0.61	0.70	0.70
CALM	0.58	0.49	0.77
DISGUST	0.44	0.39	0.65
FEAR	0.43	0.59	0.60
HAPPY	0.38	0.31	0.57
NEUTRAL	0.29	0.26	0.37
SAD	0.47	0.39	0.60
SURPRISE	0.58	0.41	0.70

Table 2 Accuracy of different Classification models

Classification Model	Accuracy (%)
SVM	48.61
MLP	46.11
CNN	63.88

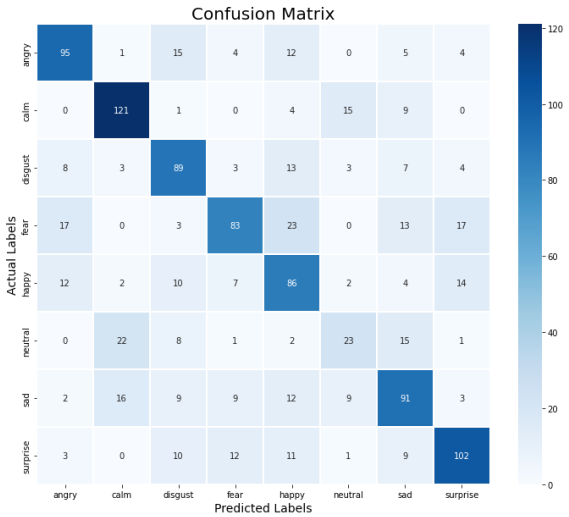

Fig.2: Confusion Matrix for CNN Model

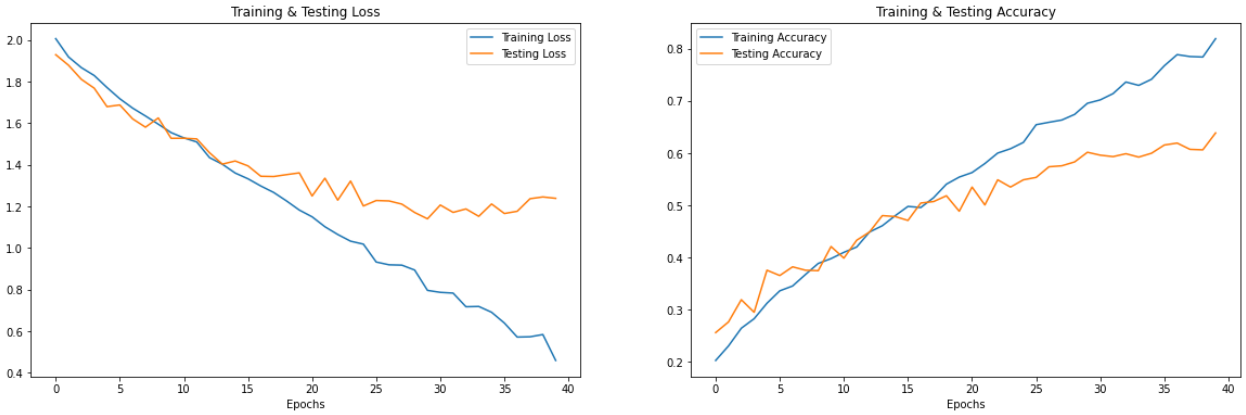

Fig.3: CNN Model Loss and Accuracy plot against epochs for Training and Testing Data

VI. Conclusion

In this project, it has been shown that CNN works better with RAVDESS Dataset compared to SVM and MLP classifiers. This machine learning method can be used to extract emotion from human speech data. This system can be useful for different fields like Call Centre for marketing or reporting, voice-based virtual assistants or chatbots etc. To improve the accuracy of this model different combinations of parameters in CNN model can be implemented.

REFERENCES:

1. Iqbal, Aseef & Barua, Kakon. (2019). A Real-time Emotion Recognition from Speech using Gradient Boosting. 1-5. 10.1109/ECACE.2019.8679271. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.
2. Chaudhary AN, Sharma AK, Dalal JY, Choukiker LE. Speech emotion recognition. J Emerg Technol Innov Res. 2015;2(4):1169-71.
3. Xu Dong an and Zhou Ruan 2021 J. Phys.: Conf. Ser. 1861 012064.
4. Logan, Beth. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. Proc. 1st Int. Symposium Music Information Retrieval.
5. B. Schuller, G. Rigoll and M. Lang, "Hidden Markov model-based speech emotion recognition," 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., Hong Kong, China, 2003, pp. II-1, doi: 10.1109/ICASSP.2003.1202279.
6. Joshi, Aastha. “Speech Emotion Recognition Using Combined Features of HMM & SVM Algorithm.” (2013).
7. Keshi Dai, Harriet J. Fell, and Joel MacAuslan. 2008. Recognizing emotion in speech using neural networks. In Proceedings of the IASTED International Conference on Telehealth/Assistive Technologies (Telehealth/AT '08). ACTA Press, USA, 31–36.
8. Steven R. Livingstone, & Frank A. Russo. (2019). RAVDESS Emotional speech audio [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DSV/256618.