Educational Administration: Theory and Practice
2024, 30(6)(s), 118-124
ISSN: 2148-2403
https://kuey.net/

Research Article

Speech Emotion Recognition Using Machine Learning

Prof. Kinjal S. Raja[1*], Prof. Disha D. Sanghani[2]

[1*]Assistant Professor, Department of Computer Engineering, Atmiya University Rajkot, Gujarat, India, kinjal.raja@atmiyauni.ac.in
[2]Assistant Professor, Department of Information Technology Shantilal Shah Engineering College Bhavnagar, Gujarat, India, ddsanghani@it.ssgec.ac.in

Citation: Prof. Kinjal S. Raja et al. (2024) Speech Emotion Recognition Using Machine Learning, *Educational Administration: Theory And Practice*, *30* (6) (s), 118-124
Doi: 10.53555/kuey.v30i6(S).5333

ARTICLEINO

ABSTRACT

Speech signals is being considered as most effective means of communication between human beings. Many researchers have found different methods or systems to identify emotions from speech signals. Here, the various features of speech are used to classify emotions. Features like pitch, tone, intensity are essential for classification. Large number of the datasets are available for speech emotion recognition. Firstly, the extraction of features from speech emotion is carried out and then another important part is classification of emotions based upon speech. Hence, different classifiers are used to classify emotions such as Happy, Sad, Anger, Surprise, Neutral, etc. Although, there are other approaches based on machine learning algorithms for identifying emotions.

Speech Emotion Recognition is a current research topic because of its wide range of applications and it became a challenge in the field of speech processing too. We have carried out a brief study on Speech Emotion Analysis along with Emotion Recognition. Speech Emotion Recognition (SER) can be defined as extraction of the emotional state of the speaker from his or her speech signal. There are few universal emotions including Neutral, Anger,. we have worked on different tools to be used in SER. SER is tough because emotions are subjective and annotating audio is challenging task.

Emotion recognition is the part of speech recognition which is gaining more popularity and need for it increases enormously. We have classified based on different type of emotions to detect from speech.

Keywords—Speech Recognition , Machine Learning , Emotion Recognition , Deep Learning

Introduction

Speech Emotion Recognition, in short can be identified as SER, is useful for the recognition of human emotions and its different states of behavior from speech. The primary goal of Speech Emotion Recognition is to develop computational models that can accurately identify and classify the emotional content expressed in spoken language.

The voice is recognized based on its tone and pitch signals. This is also the process that animals like dogs and horses are performing to be able to understand human emotion. Emotions are subjective, and audio annotation is difficult, which makes SER difficult. The aspect of speech recognition that is becoming more important and necessary is emotion recognition.

An illustration of the domains and applications of these studies is as follows:

1) *Education: An online course system has the ability to identify disinterested participants, allowing them to alter the content's style or difficulty level.*

2) *Vehicle: There is frequently an intrinsic correlation between a driver's mental condition and driving performance. As a result, these systems can be applied to enhance driving enjoyment and driving efficiency*

3) *Security: They can be used as support systems in public spaces by detecting extreme feelings such as fear and anxiety [6].*

4) *Communication: Call centers can enhance customer service by integrating the interactive voice response system with the automatic emotion recognition system.*

5) *Health: It can be beneficial for people with autism who can use portable devices to understand their own feelings and emotions and possibly adjust their social behavior accordingly [6].*

The vast majority of SER problems involve extracting key audio features like Mel-frequency Cepstral Coefficient (MFCC), Constant-Q Transform (CQT).

Emotions play a crucial role in human communication. The ability to recognize emotions in speech is essential for various applications, including human-computer interaction, virtual assistants, customer service systems, and psychological research.

II. PROPOSED METHODOLOGY

A. Datasets:

For this research work, Two sets of data have been used. The first is the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [35], and the second is the Toronto Emotional Speech Set (TESS) [34]. The 200 target words that make up TESS are all spoken after the prompt, "Say the word." Two women who portrayed seven different emotions—anger, disgust, fear, happiness, pleasant surprise, sadness, and neutral—have recorded the same thing. The dataset has 2800 data samples in total. Each of the 24 actors in RAVDESS has 60 trials. A total of eight emotions are portrayed (calm, happy, sad, angry, fearful, surprise, and disgust) in this dataset. There are a total of 1440 data samples in the dataset.

III. PROPOSED SYSTEM

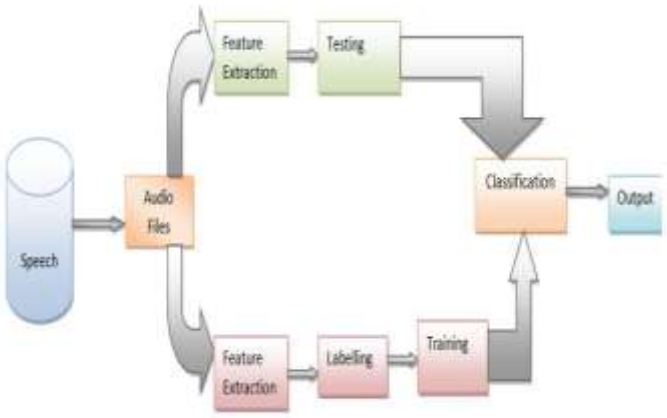

Fig.1 Block Diagram of Proposed System

A. Pre-processing:

An action taken on the speech samples prior to obtaining the signal's characteristics is known as preprocessing.Speech samples include extra details such as background noise and variations in the recording environment. At this point, the filters can be used to get rid of them.

B. Feature Extractions:

We must recognize and take out one or more significant characteristics from a speech in order to predict its emotional content.When appropriate features are extracted from voice signals, an audio dataset is made suitable.In this approach, a combination of MFCC, Log Mel-Spectrogram, Chroma, Spectral centroid, and Spectral rolloff have been extracted.

1) MFCC (Mel Frequency Cepstral Coefficients):

Any periodic component, like echoes, appears as strong peaks in the associated frequency spectrum, or Fourier spectrum, when time signals are analyzed conventionally. Applying a Fourier transform to the time signal yields this result. A spectrogram can be used to apply the Fourier Transform to acquire any cepstrum characteristic. The unique feature of MFCC is that it is measured on a Mel scale, which connects a tone's perceived frequency to its actual measured frequency. It adjusts the frequency to more closely resemble what the human ear can detect.

2) Mel Spectrogram:
The spectrogram is obtained by performing a Fast Fourier Transform on overlapping windowed slices of the signal. All that's seen here is a spectrogram that maps amplitude on a Mel scale.[2]

3) Chroma:
A Chroma vector is basically a 12-element feature vector showing how much energy of each pitch class is present in the signal .[2]

C. Feature Selections:
The process of choosing features from extracted features to eliminate redundant and unnecessary data as well as to speed up processing—a big amount of data necessitates more processing time—is known as feature selection. Within this framework MFCC includes extra data in order to reduce processing time. The values of the global features—Min, Max, Mean, Median, and Standard Deviation—are selected.These values are taken out of the MFCC feature and utilized as the MFCC feature's final features.

D. Classification Algorithms:
1) Support Vector Machines (SVM):SVM is a supervised learning algorithm that can be used for both binary and multi-class classification. It works well in high-dimensional feature spaces and is effective when there is a clear margin of separation between different classes.

2) Random Forest:Random Forest is an ensemble learning algorithm that combines the predictions of multiple decision trees. It is robust, handles non-linear relationships well, and can handle a large number of features.

3) k-Nearest Neighbors (k-NN):k-NN is a simple and intuitive algorithm that classifies samples based on the majority class of their k-nearest neighbors. It works well for small to medium-sized datasets.

4) Neural Networks:Deep learning techniques, especially neural networks, have shown great success in various classification tasks, including SER. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are commonly used for sequential data like speech.

5) Hidden Markov Models (HMM):HMMs are widely used for modeling sequential data. In SER, HMMs can capture the temporal dynamics of speech and emotion transitions.

6) Gaussian Mixture Models (GMM):GMMs are often used in combination with Universal Background Models (UBM) for speaker and emotion modeling. They model the probability distribution of the features for each emotion class.

IV. RESULTS AND DISCUSSIONS

The choice of datasets plays a significant role in the performance of SER models. Researchers often evaluate their models on multiple datasets to assess generalization capabilities. The diversity of datasets, including recordings in different languages, cultural contexts, and emotional expressions, is crucial for developing robust and generalizable models.

These are the results of the different emotions and its speech signal associated to it. The separate spectrogram graphs are also found fpr each different emotions.

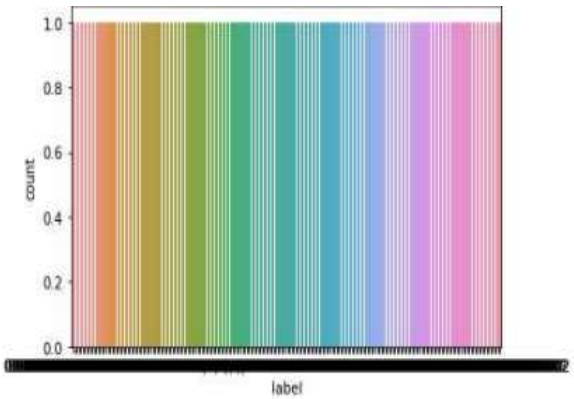

Fig.2 Exploring Data Analysis

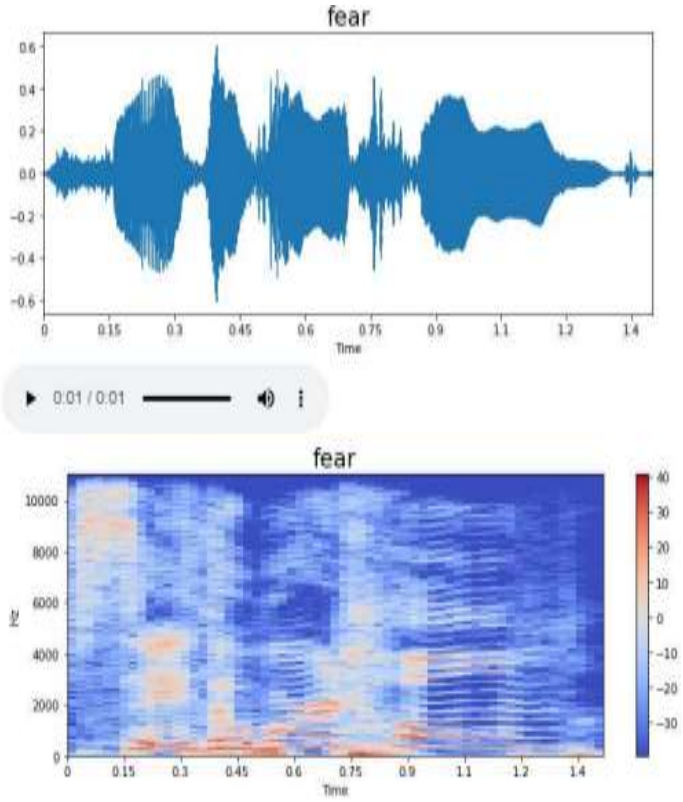

Fig.3 Specify Fear Emotions

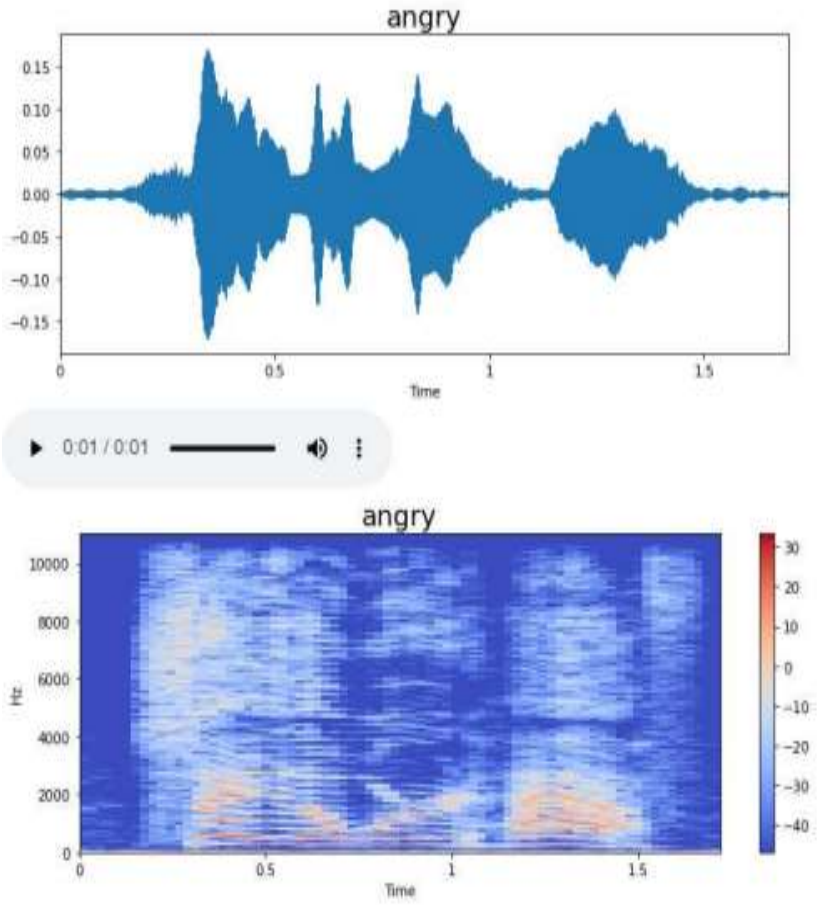

Fig.4 Specify Angry Emotions

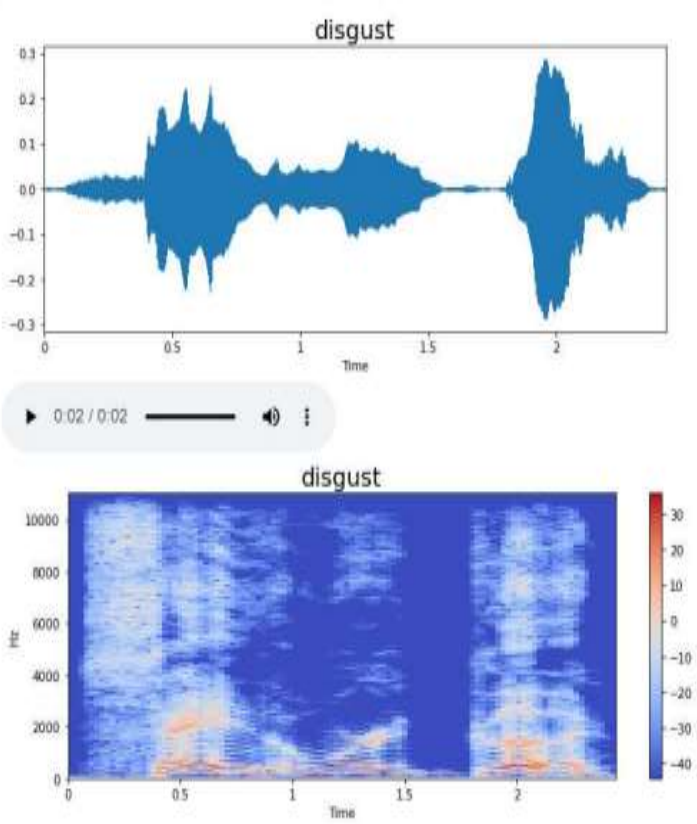

Fig.5 Specify Disgust Emotions

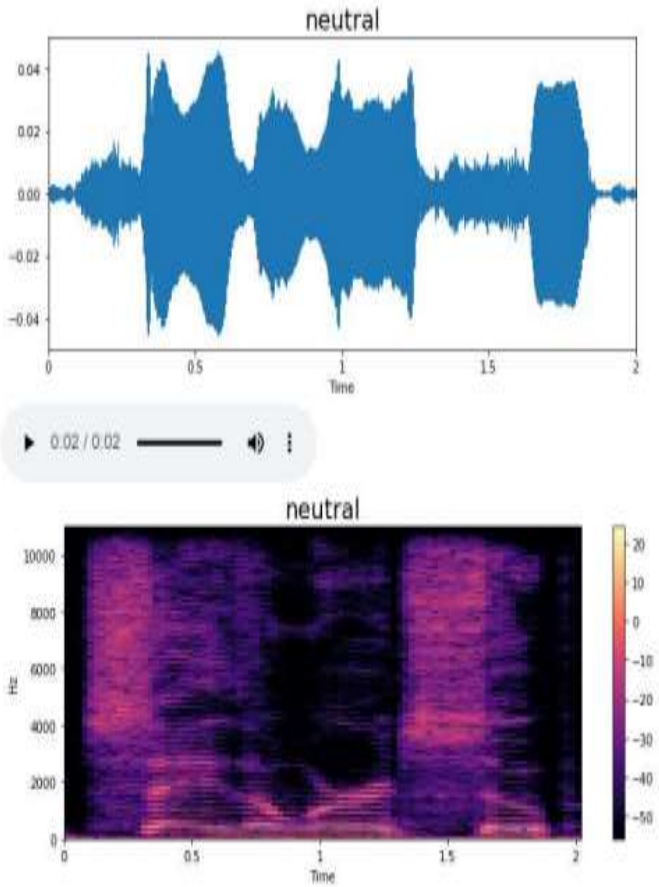

Fig.6 Specify Neutral Emotions

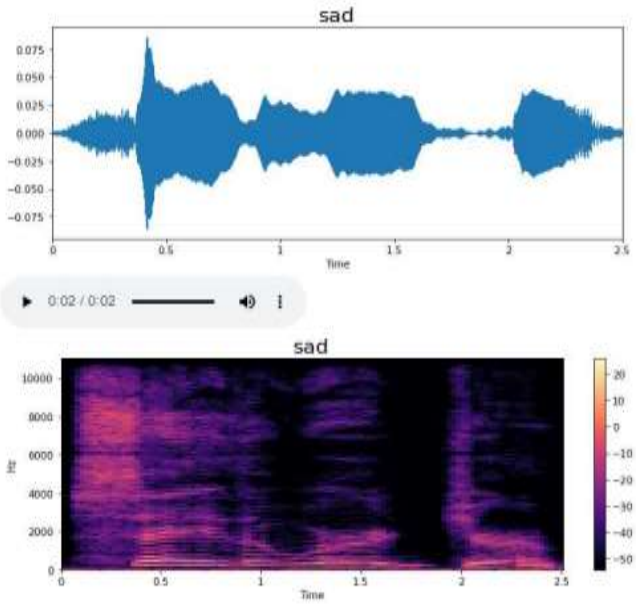

Fig.7 Specify Sad Emotions

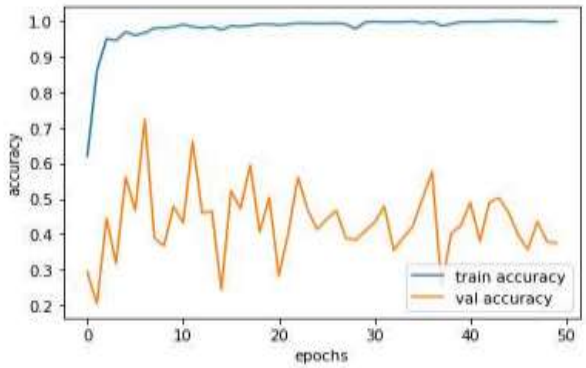

Fig.8 Accuracy Result

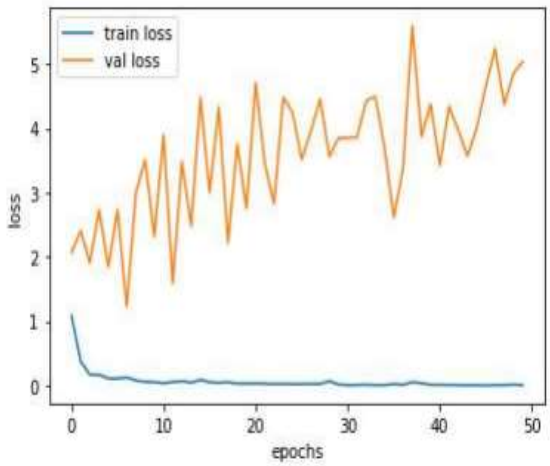

Fig.9 Loss Result

V. CONCLUSIONS

we can leverage Machine learning to obtain the underlying emotion from speech audio data and some insights on the human expression of emotion through voice. This system can be employed in a variety of setups like Call Centre for complaints or marketing, in voice-based virtual assistants or chatbots, in linguistic research, etc.Detect Forgery of Emotions from Phone Call. This research aims to improve the accuracy of current prediction model and help in call centres. The available dataset contains large number of features

which leads to over fitting of the model and reduced accuracy. The accuracy can be improved by adding pre-processing steps such as data cleaning and dimensionality reduction

References

[1] Ittichaichareon, C. (2012). Speech recognition using MFCC. ... Conference on Computer ..., 135–138. https://doi.org/10.13140/RG.2.1.2598.3208
[2] Akçay MB, Oğuz K (2020) Speech emotion recognition: emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers. Speech Commun 116:56–76. https://doi.org/10.1016/j.specom.2019.12.001
[3] https://ieeexplore.ieee.org/abstract/document/9640995/
[4] Sezgin, M. C., Gunsel, B., & Kurt, G. K. (2012). Perceptual audio features for emotion detection. EURASIP Journal on Audio, Speech, and Music Processing, 2012(1), 16. https://doi.org/10.1186/1687-4722-2012-16
[5] https://www.frontiersin.org/articles/10.3389/fcomp.2020.00014/full
[6] https://link.springer.com/article/10.1007/s40747-021-00295z https://doi.org/10.1007/s40747-021-00377-y
[7] https://www.analyticsinsight.net/speech-emotion-recognition-ser-through-machine learning/
[8] Albahri, A., Lech, M., and Cheng, E. (2016). Effect of speech compression on the automatic recognition of emotions. Int. J. Signal Process. Syst. 4, 55–61. doi: 10.12720/ijsps.4.1.55-61
[9] André, E., Rehm, M., Minker, W., and Bühler, D. (2004). “Endowing spoken language dialogue systems with emotional intelligence,” in Affective Dialogue Systems Tutorial and Research Workshop, ADS 2004, eds E. Andre, L. Dybkjaer, P. Heisterkamp, and W. Minker (Germany: Kloster Irsee), 178–187.
[10] Bachorovski, J. A., and Owren, M. J. (1995). Vocal expression of emotion: acoustic properties of speech are associated with emotional intensity and context. Psychol. Sci. 6, 219–224.
[11] Badshah, A. M., Ahmad, J., Rahim, N., and Baik, S. W. (2017). “Speech emotion recognition from spectrograms with deep convolutional neural network,” in 2017 International Conference on Platform Technology and Service (PlatCon-17) (Busan), 1–5.
[12] Bui, H. M., Lech, M., Cheng, E., Neville, K., and Burnett, I. (2017). Object recognition using deep convolutional features transformed by a recursive network structure. IEEE Access 4, 10059–10066. doi: 10.1109/ACCESS.2016.2639543
[13] ayek, H. M., Lech, M., and Cavedon, L. (2017). Evaluating deep learning architectures for speech emotion recognition. Neural Netw. 92, 60–68. doi: 10.1016/j.neunet.2017.02.013
[14] Al-Talabani, A., Sellahewa, H., &Jassim, S. A. (2015). Emotion recognition from speech: tools and challenges. Mobile Multimedia/Image Processing, Security, and Applications 2015, 9497(May 2020), 94970N.https://doi.org/10.1117/12.2191623
[15] Deep learning approaches for speech emotion recognition: State of the art and research challenges

a.