International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Intelligent Speech Emotion Classification Using Deep Learning

Harsh Suryawanshi[1]*, Anuj Zanwar*[2]*, Prathmesh Yadav*[3]*, Ashwin Ramteke*[4]

[1]Department of Electronics and Telecommunication Pune Institute of Computer Technology
[2]Department of Electronics and Telecommunication Pune Institute of Computer Technology
[3]Department of Electronics and Telecommunication Pune Institute of Computer Technology
[4]Asst. Professor Department of Electronics and Telecommunication Pune Institute of Computer Technology

ABSTRACT –

In the field of speech emotion recognition, many techniques have been used to extract emotion from signals, including many well-established speech analysis and classification techniques. In the traditional way of speech emotion recognition, the emotion recognition features are extracted from the speech signals, and then the features, which are collectively known as the selection module, are selected, and then the emotions are recognized, which is a very tedious and time-consuming process, so this paper provides an overview of the deep learning technique , which is based on a simple algorithm based on feature extraction and building a model that recognizes emotions

I. INTRODUCTION

The Speech emotion recognition (SER) is the task of recognizing the emotional aspects of speech independently of the semantic content. While humans can efficiently perform this task as a natural part of voice communication, the ability to use programmable devices to perform this automatically is still a subject of research. Call centre operators and customers, drivers, pilots, and many other users of human-machine communication. gain. Adding emotion to machines is recognized as a key factor in making them look and behave like humans. Robots that are able to understand emotions can display appropriate emotional responses and display emotional personalities.

A. Background

1) According to s. Cao et al. [1] . proposes a new approach to emotion recognition that uses a ranking support vector machine (SVM) to synthesize information. The paper addresses the challenge of binary classification in emotion recognition by proposing a method that ranks emotions based on their relevance and importance in a given context. The proposed method shows promising results in accurately identifying emotions and could have potential applications in fields such as human-computer interaction and psychology. Overall, the paper contributes to the development of more effective and contextually relevant methods for emotion recognition.

2) According to Chen et al. [2] proposes a new approach to speech emotion classification that combines deep neural networks with acoustic and lexical features. The paper addresses the challenge of accurately classifying emotions in speech signals, which has important applications in fields such as affective computing and human-robot interaction. The proposed method shows promising results in accurately identifying emotions from speech signals and outperforms previous state-of-the-art approaches. Overall, the paper contributes to the development of more effective and accurate methods for speech emotion classification using a combination of acoustic, lexical, and deep neural network features

3) According to Yang & Lugger.al. [3] The paper presents a technical implementation of a speech emotion classification system that uses a combination of acoustic features, machine learning algorithms, and a web-based user interface. The paper addresses the challenge of creating an effective and user-friendly system for emotion recognition from speech signals, which has important applications in fields such as psychology, social robotics, and virtual assistants. The proposed system shows promising results in accurately identifying emotions from speech signals, and the user interface provides a user-friendly way to interact with the system. Overall, the paper contributes to the development of more accessible and user-friendly methods for speech emotion classification with potential practical applications in various domains. The Yang and Lugger paper uses several machine learning algorithms for speech emotion classification, including support vector machines (SVM), k-nearest neighbors (KNN), decision trees (DT), and random forests (RF). These algorithms are trained on a combination of acoustic features extracted from speech signals, including pitch, energy, and spectral features. The paper also employs feature selection techniques to identify the most relevant features for classification. The proposed system uses a combination of these algorithms to accurately classify emotions from speech signals and provides a user-friendly web-based interface for interacting with the system.

4) According to Albornoz et al. [4] (2018) presents a method for classifying emotions in speech using deep neural networks. The authors explore the use of two different models: a convolutional neural network (CNN) and a long short-term memory (LSTM) network.The authors used the

Berlin Emotional Speech Dataset (EmoDB) to train and test their models. This dataset consists of recordings of actors speaking six different emotions: anger, boredom, disgust, fear, happiness, and sadness.To preprocess the data, the authors extracted Mel frequency cepstral coefficients (MFCCs) from each audio recording, which are commonly used features for speech analysis. They then used these features to train their CNN and LSTM models. The CNN model consists of three convolutional layers followed by two fully connected layers. The LSTM model consists of one LSTM layer followed by two fully connected layers. Both models were trained using stochastic gradient descent with backpropagation. The authors evaluated the performance of their models using accuracy, precision, recall, and F1-score metrics. They found that both models achieved high accuracy in classifying emotions in speech, with the LSTM model slightly outperforming the CNN model. Overall, the paper demonstrates the effectiveness of deep neural networks for speech emotion classification and provides a useful technical implementation for future researchers in this field.

5) According to Albornoz et al. [5] (2014) proposed a paper titled "Spectral feature selection for emotion recognition from speech using Gaussian mixture models." The paper presents a methodology for recognizing emotions from speech by selecting spectral features and using Gaussian mixture models (GMMs) to classify emotions.The authors analyzed several spectral features such as mel-frequency cepstral coefficients (MFCCs), spectral entropy, spectral roll-off, and spectral flux. They evaluated the performance of these features in recognizing emotions using two datasets, namely the Berlin Emotional Speech Database and the Emotional Prosody Speech and Transcripts dataset

II. PROPOSED METHODOLOGY

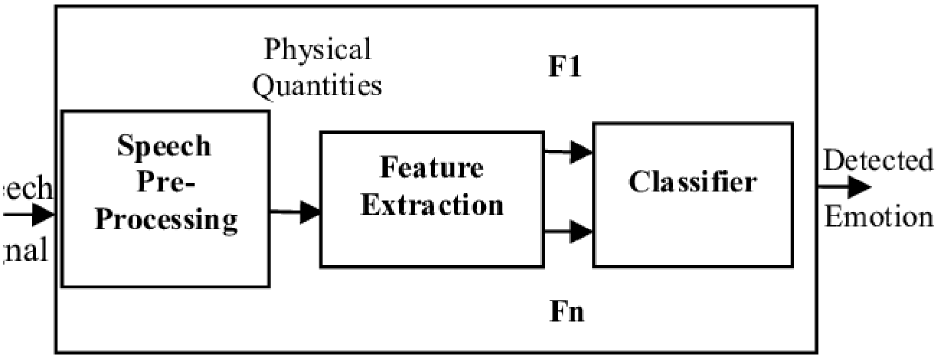
1) In Data preprocessing: The speech signal is first preprocessed by segmenting it into frames of a fixed duration, typically 20-30 milliseconds. Each frame is then transformed into a Mel-frequency cepstral coefficient (MFCC) feature vector.

2) Feature extraction: MFCC is a widely used feature extraction technique that captures the spectral characteristics of speech signals. It involves taking the logarithm of the short-term power spectrum of the speech signal, followed by a Discrete Cosine Transform (DCT) to obtain a set of coefficients that represent the spectral envelope of the signal

3) LSTM Model Architecture: LSTM (Long Short-Term Memory) is a type of recurrent neural network (RNN) that is well-suited for modeling sequential data. The LSTM model takes in a sequence of MFCC feature vectors as input and outputs a label indicating the emotion present in the input sequence.

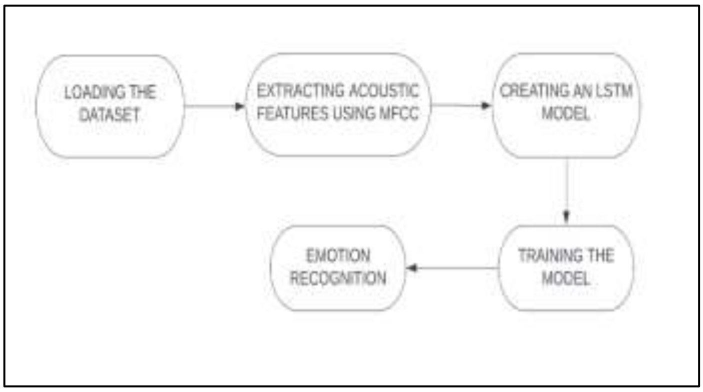
III. SYSTEM ARCHITECTURE

Fig 1



The Fig. 1 shows complete architecture of Speech Emotion Classification. By extracting Features and model as classifier.

Fig 2



The Fig.2 shows the flow of the process adapted for classification of input speech signal based on input signal.

LSTM:-

1) One of the key advantages of LSTM models over traditional RNNs is their ability to model long-term dependencies in sequential data.
2) LSTM models are highly flexible in handling input and output sequences of varying lengths. This makes them well-suited for modeling a wide range of sequential data, such as speech signals, text, and time-series data.

1) 2 Datasets will be used , first dataset of different samples (TESS) dataset which consists of 2800 different types of audio samples.
2) Second dataset is (RAVDESS) The dataset contains 7356 audio and video recordings of 24 professional actors (12 male and 12 female) speaking and singing in different emotional states.

IV. CONCLUSION AND FUTURE SCOPE

A. Conclusion

In this research, deep learning classification algorithms like MFCC, LSTM were studied. We have proposed a model which identifies the speech's emotion. MFCC extract features based on variations of specified parameters. The model helps determine emotion based on 7 classification determined and makes a speech fall under the specified category.

B. Future Scope

The speech emotion recognition is a very interesting topic and there is much more to discover in this field, in our model future work will include improving the accuracy of the model to achieve better results, we can also train the model to provide results from speech that lasts longer as in this model, we are only able to recognize the emotion for a short period of time in the future we will be able to load a longer sample data set and the model will classify different emotions in different time periods. His future work may also involve recording real-time data using a microphone, so there is no need to load the dataset, we just train the model and then the data can be recorded to provide the emotion of that person's voice.

ACKNOWLEDGMENT

We take this opportunity to express our sincere thanks to our guide respected Mr. Ashwin Ramteke sir for their support and encouragement throughout the completion of this project. We are grateful to him for his

REFERENCES

[1] s. Cao et al.[1] proposed a ranking SVM method for synthesize information about emotion recognition to solve the problem of binary classification.

[2] Chen et al. [2] aimed to improve speech emotion recognition in speaker-independent with three level speech emotion recognition method.

[3] Yang & Lugger [3] presented a novel set of harmony features for speech emotion recognition. These features are relying on psychoacoustic perception.

[4] Albornoz et al. [4] investigate a new spectral feature in order to determine emotions and to characterize groups.

[5] Lee et al. [5] represent a hierarchical computational structure to identify emotions. Lee et al. [11-12] proposed hierarchical structure for binary decision tree in emotion recognition fields.

[6] Yeh et al. proposed a segment based method for recognition of emotion in Mandarin speech

[7] El Ayadi et al. proposed a Gaussian mixture vector autoregressive (GMVAR) approach, which is mixture of GMM with vector autoregressive for classification problem of speech emotion recognition.

[8] B. W. Schuller, ‘‘Speech emotion recognition: Two decades in a nut shell, benchmarks, and ongoing trends,’’ Commun. ACM, vol. 61, no. 5, pp. 90– 99, 2018.