International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 03 | March - 2023 Impact Factor: 7.185 ISSN: 2582-3930

SPEECH EMOTION RECOGNITION USING CNN-LSTM

Ms Gayathri R [1], Arun Kumar B[2], Inbanathan S[3], Karthick S[4]

[1]Assistant Professor (Sr.Gr), Department of Electronics and Communication Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India.

[2]Final Year Student, Electronics and Communication Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India.

[3]Final Year Student, Electronics and Communication Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India.

[4]Final Year Student, Electronics and Communication Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India.

Abstract -Speech emotion recognition is a rapidly growing field of research that aims to automatically identify emotions from speech signals. This paper presents a speech emotion recognition using machine learning techniques. The study begins by providing an overview of the various approaches used in speech emotion recognition, including feature extraction, feature selection, and classification. These selected features like pitch, MFCC are compared with the existing datasets in databases. and baased on the features the audios are classified using CNN LSTM algorithm. This model is trained in the free environments like collab using Python, and for User interface and HTML, CSS is used.

Key Words: Speech Emotion, MelFrequency Cepstral Coefficient, CNN, LSTM

I. INTRODUCTION

Speech emotion recognition is a rapidly growing field of research that aims to automatically identify emotions from speech signals. This paper presents a speech emotion recognition using machine learning techniques.The study begins by providing an overview of the various approaches used in speech emotion recognition, including feature extraction, feature selection, and classification. It then discusses the challenges associated with speech emotion recognition, such as the variability of emotional expression across different speakers and cultures also presents a comprehensive review of recent research in speech emotion recognition using machine learning techniques, including support vector machines, artificial neural networks, and deep learning approaches such as convolutional neural networks and recurrent neural networks. The strengths and weaknesses of each approach are discussed, along with their respective applications in speech emotion recognition.

II. LITERATURE SURVEY

[1], Pavol Harar presented a method that achieved 96.97% accuracy on testing and 69.55% on file prediction. In this method, Deep Neural Network (DNN) architecture with convolutional, pooling and fully connected layers was used for emotion recognition.

[2] Supriya B. Jagtap, Presented a system to detect seven emotions that are happiness, Anger, Boredom, Sadness, Surprise, Fear, and Neutral emotions. The study presents frequency information contained in speech signal are reduced into small numbers of coefficients using MFCC. This study also presents that the accuracy of the system depends on the database used for training.

[3], Shumin presented methods based on LSTM-RNN models. This method has achieved 96.67% accuracy in case of angry emotion, 100% accuracy in case of sad emotion and for natural 86.67% accuracy is achieved. A literature survey shows that the MFCC feature is the most popular choice to identify emotions.

[4] K. Juglan et al. A system for acknowledging emotions in Punjabi dialect was proposed in 2018.. A person's enthusiastic state just as its semantic arrangement can be resolved from their discourse. In this paper, the sample of 120 grown-ups (58 males and 62 females) was taken for the exploratory investigation. Discourse test comprised of words tokens. In every speaker rehashed five Punjabi sentences with five distinct emotions to extricate pitch and intensity. Recording and naming of word tokens were finished with help of PRAAT programming. According to estimations, the creator believed that the estimation of the pitch for typical emotions was highest when happy occurred and lowest when sad occurred.

[5] The significance of speech emotion datasets and features, noise reduction, and various classification approaches, including SVM and HMM, are all covered in a brief review Basu et al. published in 2020. The research's strength is the discovery of a number of features associated with speech emotion recognition; however, its flaw is the investigation of more contemporary methods' leak and the brief mention of

convolutional and recurrent neural networks as a deep learning technique..

III. METHODOLOGY

Speech emotion recognition using machine learning involves several steps, including data collection and pre-processing, feature extraction, training a model, and evaluating the model's performance. Here is a high-level overview of the methodology:

Data Collection and Pre-processing: The first step is to collect a dataset of speech samples that represent different emotions. The dataset should be pre-processed to remove noise and other artifacts that could interfere with the analysis. In many speech analysis systems, the choice of whether to classify a particular speech signal as a voiced speech section or a quiet section must be made. Non-speech portions might supply extraneous information and provide a challenge when they are incorporated into the teaching or testing process. This step may also involve data augmentation techniques to increase the size and diversity of the dataset.

Feature Extraction: The next step is to extract features from the pre-processed speech samples. Features are numerical representations of the speech signals that capture the relevant information for the emotion recognition task. Commonly used features for speech emotion recognition include Mel-Frequency Cepstral Coefficients (MFCCs), Linear Predictive Coding (LPC) coefficients, and pitch and energy measures.

Model Training: Once the features have been extracted, the next step is to train a machine learning model on the dataset. Commonly used models for speech emotion recognition include Support Vector Machines (SVM), Random Forests, and Neural Networks.

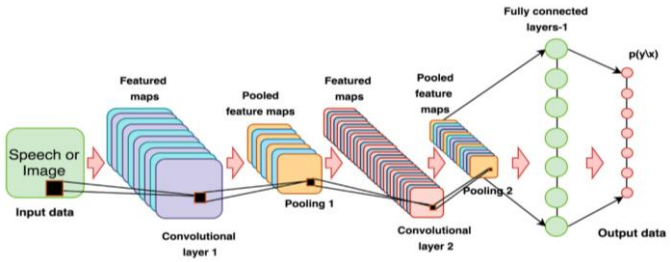
Model Evaluation: After the model has been trained, it is evaluated on a separate test set to measure its performance.

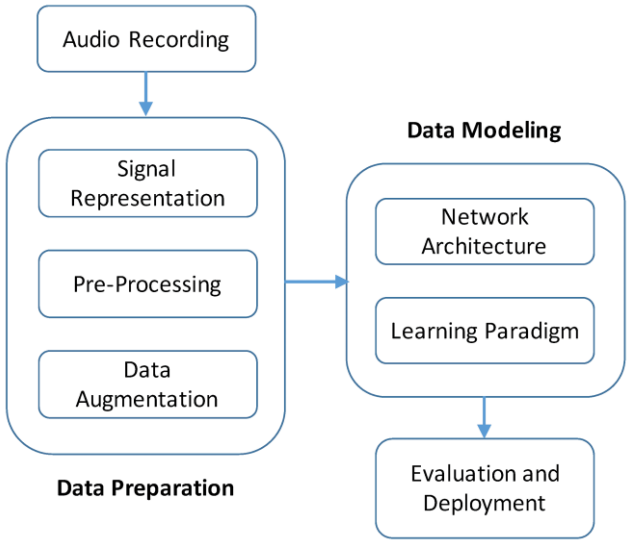
Model Deployment: Once the model has been trained and evaluated, it can be deployed in a real-world application to recognize emotions in speech signals.This is a high-level overview of the methodology for speech emotion recognition using machine learning.

The specific techniques and algorithms used may vary depending on the specific application and dataset.In many speech analysis systems, the choice of whether to classify a particular speech signal as a voiced speech section or a quiet section must be made. Non-speech portions might supply extraneous information and provide a challenge when they are incorporated into the teaching or testing process. Since the speech signal segment's signal energy value was higher than the non-speech signal segment's, an absolute integral value (IAV) reflecting the energy value was applied. The short sample size of the dataset makes the heavily parameterized neural network model in this article susceptible to overfitting.

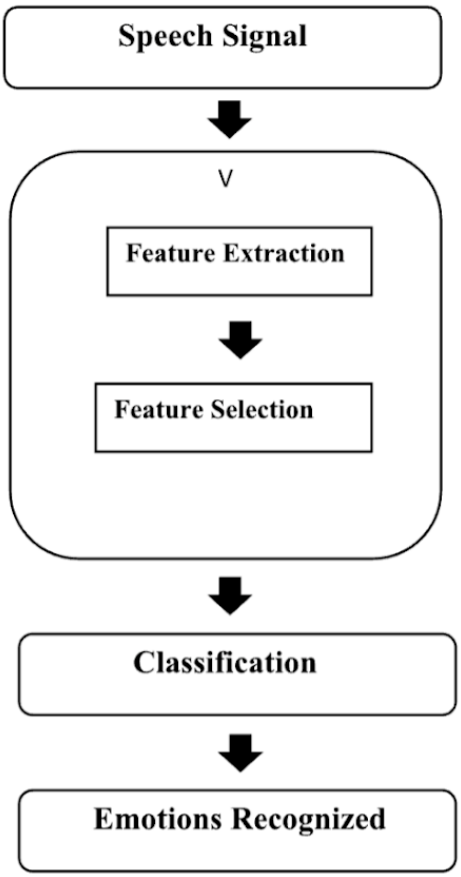
As a result, the data augmentation mechanism is included into our design. Nonetheless, it would be quite challenging to produce more real samples. Too little noise will be meaningless, and too much noise would make it impossible for the network to learn from the training data. We need a proper mix of noise. A channel model known as additive white Gaussian noise (AWGN) assumes that the only interference with communication is the linear addition of broadband or white noise with a constant spectral density (measured in watts per hertz bandwidth) and amplitude Gaussian distribution. Fading, frequency selectivity, interference, nonlinearity, and dispersion are not taken into account by this model. Additive White Gaussian Noise will be used (AWGN). The noise is additive because we are adding it to the source audio signal, gaussian because it will be sampled from a normal distribution and have a zero-time average (zero mean), and white because it will add power to the audio signal uniformly across the frequency distribution after the whitening transformation

IV. CNN ARCHITECTURE





V. BLOCK DIAGRAM



VI. SCOPE OF THE PROJECT

A Speech Emotion Recognition (SER) system is created to analyse human emotion using speech as an input for a dataset that contains audio files of various actors. Using the CNN Algorithm and the RAVDESS dataset, the system extracts, characterises, and recognises information about the speaker's emotions for both male and female speakers. The only language for which the system can accept real-time input speech data is English. As multiple datasets are being used, it is also expected to recognise more than 4 emotions. The use of the word "responsible" in the title of this post is a reference to the fact that the word "responsible" is a term that is used in the context of the United States.

VII. RESULTS

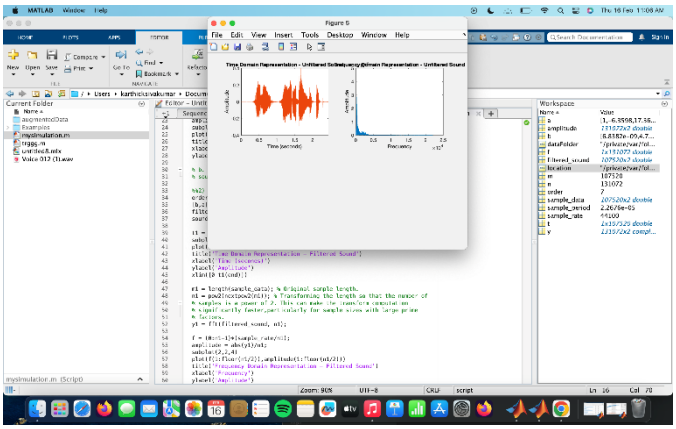


Figure 1: Preprocessing of a audio signal

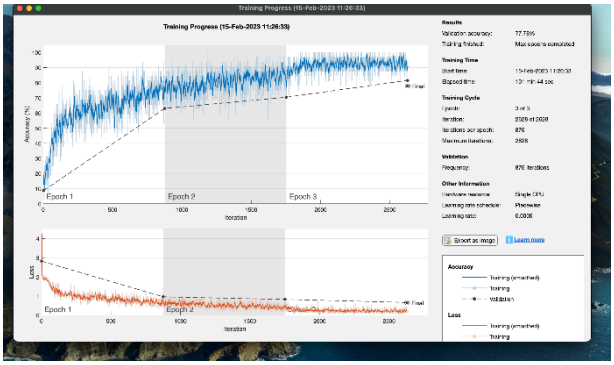


Figure 2: Training of a CNN model with Accuracy and Loss percentage with respect to the epoch

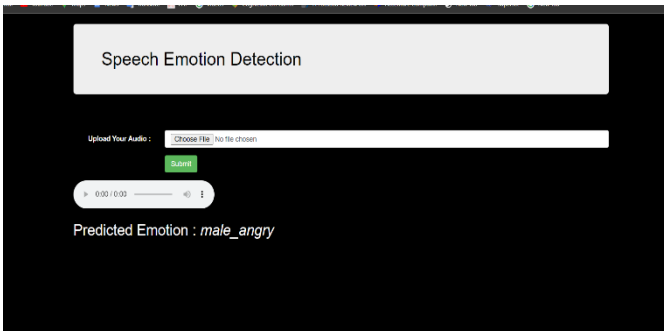


Figure 3: Predicted emotion for a given Audio Signal

VI. CONCLUSION

We developed several models before arriving at the most effective CNN model for the emotion distinction task. We were able to improve the accuracy of the model by about 80%. With more data, our model would have performed better. Additionally, our model did a fantastic job of differentiating between a male and female voice. Our project can be expanded to integrate with the robot to help it understand the mood of the corresponding human, which will help it have a better conversation. It can also be integrated with 73 different music applications to recommend songs to its users based on their emotions, and it can be used in various online shopping applications like Amazon to enhance the product recommendations for its users.

REFERENCES

[1]. Meena, S. Divya, and Loganathan Agilandeeswari. "An efficient framework for animal breeds classification using semi-supervised learning and multi-part convolutional neural network (MP-ORB). IEEE Access 7 (2019): 151783-151802.

[2]. Munian, Yuvaraj, Antonio Martinez-Molina, and Miltiadis Alamaniotis. "Intelligent system for detection of wild animals using HOG and ORB in automobile applications. 2020 11th International Conference on Information, Intelligence, Systems and Applications.

[3]. Ranparia, Devsmit, et al. "Machine learning-based acoustic repellent system for protecting crops against wild animal attacks." *2020 IEEE 15th International Conference on Industrial and Information Systems (ICIIS)*. IEEE, 2020.

[4]. Gobhinath, S., et al. "Smart irrigation with field protection and crop health monitoring system using autonomous rover." 2019 5th International Conference on Advanced Computing & Communication Systems (ICACCS).

[5]. El Abbadi, Nidhal K., and Elham Mohammed Thabit A. Alsaadi. "An automated vertebrate animals classification using deep convolution neural networks." 2020 International Conference on Computer Science and Software Engineering (CSASE). IEEE, 2020.

[6]. Giordano, Stefano, et al. "IoT solutions for crop protection against wild animal attacks." 2018 IEEE international conference on Environmental Engineering (EE). IEEE, 2018.