IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

Emotion Recognition and Analysis from Speech

Manjunatha Kumar B H, Sanjana Reddy B, Sumedha M N, Varshini S, Vedika Medha Raj,
Department of CSE
SJC Institute of Technology Chikkaballapura, India

Abstract: *This paper presents a novel approach for emotion recognition and analysis from speech using convolutional neural networks (CNNs). The proposed system involves deriving the characteristics from the audio signals using Mel Frequency Cepstral Coefficient (MFCC) and Spectrogram based representation. These extracted features will be fed into the CNN model which will be trained to classify the different input speech signals into different emotion classes. The results demonstrate the effectiveness of the proposed system in emotion analysis and recognition from speech using CNN.*

Keywords: Deep learning, convolutional neural network algorithm, Mel Frequency Cepstral Coefficient, Spectrogram, Human-Human Interaction(HHI), Human-Robot Interaction(HRI).

I. INTRODUCTION

Artificial intelligence frequently refers to human behavior as a reference. Effective communication depends heavily on emotions, and people often react and respond to their communication partner's emotions more skillfully. It has long been an aim to create AI systems that can recognize and communicate emotions.

However, as Human-Robot Interaction (HRI) technologies like voice assistants and chatbots for customer service were introduced, researchers started to create dialogue systems with empathy to enhance the overall HRI experience. A more natural HRI experience can be achieved by machines being able to recognize and respond to messages with the aid of sentiment analysis of human-human interaction (HHI).This method can aid AI in comprehending not only what people are saying, but also how they are saying it, leading to a more human-like engagement.

Convolutional Neural Networks (CNNs) are used in the proposed system to classify and extract features from speech samples. The uttered words' sentiment is first classified using the retrieved features, after which a graph representing the emotion picked up from the input voice is created. To create a more human-like interaction, the system seeks to comprehend not just what the user says but also how they say it.

In this project, we will create a sentiment analysis system utilizing real-world data. An eight-point scale of emotions can be used to express sentiment, a fundamental measure. Variability and noise, which are frequently present in real-world data, must be handled by the suggested system. The study's findings show how the suggested method for extracting sentiment from speech works well and how it could enhance HRI in general.

The structure of the rest of this paper is as follows: Section II provides an overview about the literature survey that was done. Section III provides an overview of related work in sentiment analysis and emotion recognition from speech. Section IV describes the methodology used for sentiment analysis in this study. Section V provides an overview about the workflow of proposed system. Section VI presents the results of the experiments and discussions. Finally, the paper concludes with future research directions in Section VII.

II. LITERATURE SURVEY

"Speech Emotion Recognition using Support Vector Machine" by Manas Jain *et.al.,* published in the year 2020. This research aims to identify speech using characteristics like energy, pitch, MFCC coefficients, LPCC coefficients, and speaker rate into four emotional categories: sadness, anger, fear, and happiness. The LDC and UGA databases provided samples. One against All (OAA) and Gender Dependent Classification were employed as the two classification procedures with Support Vector Machine (SVM) as the classifier. The LPCC and MFCC algorithms, as well as the two classification techniques, were also compared.

"Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations" by Bagus Tris Atmaja and Akira Sasou published in the year 2022. Using modern self-supervised learning models, comprehend

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

sentiment and emotion in evaluated sentiment analysis and emotion identification from speech. Three potential universal model sizes for sentiment analysis and emotion recognition from voice were tested using recent self-supervised learning models. For four separate tasks, three different sizes of universal models were examined. Based on weighted and unweighted accuracy ratings, two classes of sentiment analysis produced the best results. The models, however, had trouble with tasks that required a bigger number of classes for sentiment analysis and emotion recognition. The performance decline seen in several tasks may have been influenced by the uneven nature of the datasets. Comparing the performance of the binary classification utilising unimodal audio analysis to earlier multimodal fusion techniques, it was competitive.

"Dialogue RNN: An Attentive RNN for Emotion Detection in Conversations" by Navonil Majumder *et.al.,* published in the year 2019. For many applications, such as opinion mining over chat history, social media threads, disputes, argumentation mining, analysing customer feedback in live discussions, and so forth, emotion identification in talks is a prerequisite step. At the moment, systems do not take into account the speaker of each utterance and treat the participants in the conversation individually. In this paper, they proposed a novel approach based on recurrent neural networks that utilises the knowledge about the individual party states during the conversation for mood classification. On two different datasets, this model significantly outperforms the state-of-the-art.

"Multimodal and Multi-view Models for Emotion Recognition" by Gustavo Aguilar *et.al.,* published in the year 2019. Studies on emotion recognition (ER) demonstrate that more reliable and accurate models are produced when lexical and auditory data are combined. The majority of studies concentrate on contexts where in training and evaluation, both modes are available. Getting ASR output, however, may be a bottleneck in a deployment pipeline because of computational complexity issues or privacy-related restrictions. They investigated how to effectively combine acoustic and lexical modalities during training while still producing a deployable acoustic model that excludes lexical inputs in order to meet this issue. In order to determine the extent of the advantages that lexical information can offer, they first conducted experiments using multimodal models and two attention mechanisms. Then, using a contrastive loss function, they framed the task as a multi-view learning problem to integrate semantic data from a multimodal model into our acoustic-only network. On the USC-IEMOCAP dataset reported on lexical and acoustic information, our multimodal model performs better than the previous state of the art. Our multi-view-trained acoustic network outperforms models that have only been trained using acoustic features significantly.

"Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances" published by Soujanya Poria *et.al.,* in the year 2019. Emotion is ingrained in humans, thus human-like artificial intelligence (AI) must be able to comprehend it. Emotion recognition in conversation (ERC) is gaining traction as a new area of study in natural language processing (NLP) due to its ability to mine opinions from the vast amount of publicly accessible conversational data in platforms like Facebook, Youtube, Reddit, Twitter, and others. Additionally, it may be used in healthcare systems (as a psychological analysis tool), education (to recognize student annoyance), and other fields. ERC is also crucial for creating emotion-aware conversations that call for a comprehension of the user's feelings. Conversational emotion-recognition systems must be efficient and scalable in order to satisfy these needs. However, due to a number of difficulties in the research process, it is a difficult problem to resolve. In this essay, they discussed these difficulties and clarified the most current findings in this area of study. They also go over these approaches' shortcomings and the reasons why they fall short in addressing the difficulties of ERC research.

III. RELATED WORK

Emotion recognition and analysis from speech is a growing field of research that aims to automatically recognize and understand the emotional state of a speaker based on their speech signals. Here are some related works in the field of emotion recognition and analysis from speech:

"Deep Emotion Recognition from Speech using Transfer Learning" by Sahu et al. (2021) proposed a deep neural network-based approach for emotion recognition from speech using transfer learning. The proposed model achieved state-of-the-art performance on the IEMOCAP dataset, which consists of acted and improvised dialogues between two actors.

"Emotion Recognition in Speech using Gaussian Mixture Model and Deep Belief Network" by Zhang et al. (2017) proposed a hybrid approach that combines a Gaussian mixture model and a deep belief network for emotion

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

recognition from speech. The proposed model achieved high accuracy in recognizing four basic emotions: happiness, sadness, anger, and neutral.

"Speech Emotion Recognition Based on Joint Learning of Gaussian Mixture Model and Convolutional Neural Network" by Wang et al. (2018) proposed a joint learning framework that combines a Gaussian mixture model and a convolutional neural network for speech emotion recognition. The proposed model achieved high accuracy on the IEMOCAP dataset and the MSP-IMPROV dataset, which consists of improvised monologues.

"Emotion Recognition from Speech using Multi-task Learning with Deep Recurrent Neural Networks" by Zhang et al. (2018) proposed a multi-task learning approach that combines emotion recognition with speaker verification using deep recurrent neural networks. The proposed model achieved high accuracy in recognizing three basic emotions: happiness, sadness, and anger.

"Emotion Recognition from Speech using Convolutional Neural Networks with Multi-Task Learning" by Amiriparian et al. (2017) proposed a convolutional neural network-based approach for emotion recognition from speech using multi-task learning. The proposed model achieved high accuracy in recognizing four basic emotions: happiness, sadness, anger, and neutral, on the Emo- DB dataset.

IV. PROPOSED METHODOLOGY

Emotion plays a major role in human beings. Emotions can be predicted by people communicating with each other. There are different kinds of emotions when compared from one person to another. The emotion recognition and analysis from speech application is executed by using Convolution Neural Network.

The architecture of the proposed model can be depicted as follows:

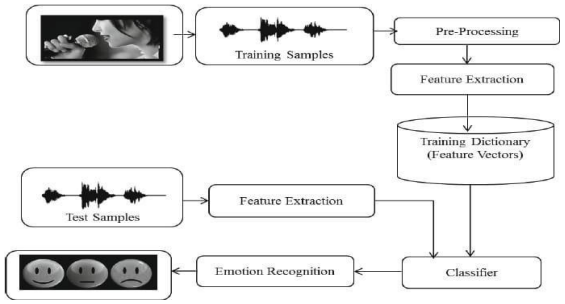

Figure 1.1 Architecture of Proposed Model

The proposed model and its methodologies are illustrated by performing these steps:

Description of Dataset

The data set holds eight different types of emotions audio files. These files are taken from Kaggle.com. There are total 1,440 datasets used for this proposed model. The eight different types of emotion includes neutral, calm, happy, sad, angry, fearful, disgust, surprised.

Feature Extraction

The first step is to collect input file which is in wave format and referred as raw data. The next step is data pre-processing which involves data cleaning, data noise reduction, data reduction, data transformation with the help of batch normalization and data validation and provide us the usable format for feature extraction. The final step is feature extraction which is process of reducing the redundant data from the data set. We use Mel-Frequency Cepstral Coefficients (MFCC) for extracting features from the audio signal. And Librosa is a python package which helps to visualize the audio signal and also do the feature extraction in it using different signal processing techniques.

MFCC: This method of feature extraction is most widely used in feature extraction process. This method uses the CNN algorithm to train the model to extract features from the audio stream and classify the features that are produced.

Convolution Neural Network (CNN) Training:

A Convolution Neural Network (CNN) is a type of deep learning neural network. It is an architecture which consists of multiple layers and are mostly used for image recognition and classification. There are four layers in CNN model:

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

- Input Layer: The input layer in CNN model incorporates spectrogram. It is a visual representation of frequencies of given audio signal with time. The representation of spectrogram is plotted where one axis represents frequencies and other axis represents the time. The colours represent magnitude of observed frequency at a particular time. Bright colour shows strong frequencies.
- Convolution Layer: This layer is occasionally know as feature extraction layer because it extracts the features from the input provided. The input is passed through the set of convolution filters that is kernels, each of which activates certain features from the image. In convolution process, each of K kernels sliding across the input region, performing element wise product and summing up which we will obtain an output of 2-D activation map form from a 3-D matrix. This layer holds ReLU (Rectified Linear Unit) which is an activation function used for mapping negative values to zero and maintaining positive values and the output will be feature map.
- Max Pooling Layer: The feature maps is passed to this layer which includes the down-sampling of the features with the goal of reducing number of parameters or decreasing the dimensionality of the features without loss of information that is required for learning.
- Fully Connected Layer (FC): A fully connected layer includes weights, biases along with neurons It is used to attach each neurons with the other neurons. This layer is always placed at the end of the network. It is common to use one or two FC layers. It uses classifiers to get the final output.
- Output Layer: This layer holds the final output of the network.

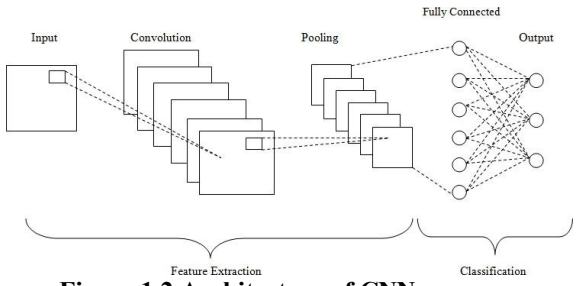

Figure 1.2 Architecture of CNN

Fig 1.2 depicts the architecture of CNN from input to output.

Evaluation Method

For validating our proposed methodology by leaving aside a portion of the data as a test set. This can be achieved by Leave-one-Speaker-out cross validation (LOSOCV) which is a speaker-independent test-runs. After the test-runs , it will classify the emotion and returns the output in the form of graph.

Pseudocode for prediction model:

1. Define a function for prediction model
2. Load a trained emotion recognition model
3. Preprocess the features if required (e.g., scaling, normalization).
4. Feed the preprocessed features to the emotion recognition model.
5. Obtain the predicted emotion label.
6. Return the predicted emotion label.
7. Load an audio file.
8. Extract features from the audio file using the feature extraction function.
9. Call the prediction model function with the extracted features as input.
10. Receive the predicted emotion label.

Pseudocode for training model:

1. Define a function for training model.
2. Apply feature extraction to the training data to obtain the feature vectors.
3. Preprocess the feature vectors if required (e.g., scaling, normalization).

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

4. Split the feature vectors and labels into training and validation sets.
5. Define an emotion recognition model architecture.
6. Initialize the emotion recognition model with the defined architecture.
7. Train the model using the training set of feature vectors and labels.
8. Evaluate the model performance on the validation set.
9. Return the trained emotion recognition model.
10. Load the training data and labels.
11. Call the training function with the training data and labels as input.
12. Receive the trained emotion recognition model.

V. WORKFLOW OF PROPOSED SYSTEM

With collected data pre-processing has been done. Once it is done feature extraction and feature selection has to be performed. Further model is trained and tested, then the validation process takes place.

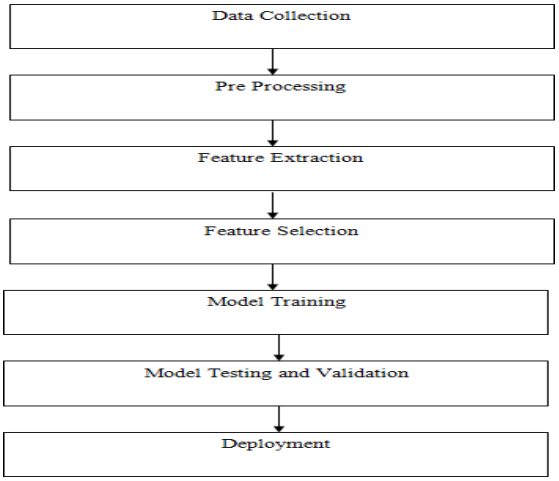

Figure 1.3 Workflow of Proposed System

VI. IMPLEMENTATION

The implementation of the proposed work can be explained as follows:

Data Collection

- The data set includes audio files representing eight distinct emotional states. These files were collected from the website Kaggle.com. For this proposed model, a total of 1,440 datasets are used.

Preprocessing

- Using the librosa.load() method, which turns the audio signal into a time series array, the audio files are loaded.
- Using librosa.resample(), the signal is resampled to a fixed rate while preserving the audio's sample rate.
- Using librosa.feature.mfcc(), the Mel- frequency cepstral coefficients (MFCCs) are calculated from the resampled signal.
- In order to create a representative feature vector for each audio file, the computed MFCCs are averaged across the time axis.

Feature Extraction:

- The librosa.feature.mfcc() function computes the MFCCs for each audio file, producing a matrix of MFCC coefficients.
- The coefficients along the time axis are then averaged using np.mean() to reshape or transform the MFCC matrix into a 1- dimensional feature vector.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

- The final feature vector is saved in a DataFrame and represents the features that were extracted from each audio file.
- Every audio file is subjected to the feature extraction procedure, producing a DataFrame that contains the feature vectors for the entire dataset.

Feature Selection:

- The audio samples are subsequently represented by the computed MFCCs as features. The code creates a one-dimensional feature vector by taking the mean of the MFCCs along the 0-axis using np.mean.
- Using the pd.DataFrame and stack functions, the feature vector is then further processed by converting it into a dataframe and stacking the values. To restructure the feature vector into an appropriate format for feeding it into the model, this step is required.
- To make the feature vector compatible with the CNN model, a third dimension is added after the feature vector is enlarged to two dimensions using np.expand_dims.

Model Training:

- The fit function is used to feed the preprocessed data into the model for training. The built-in model, training data, and predetermined batch size are used to perform the training. How many times the complete training dataset is run through the model depends on the number of epochs.
- The model modifies its weights based on the estimated loss and the optimisation technique (RMSprop) during training. The goal is to reduce the loss and raise the model's ability to correctly identify the appropriate emotion from audio samples.
- The model's weights are stored in a file for later usage after training.

Model Testing and Validation:

- A distinct set of audio samples is created in order to evaluate the model. Similar to how the training data are treated, these samples are also processed.
- The stored weights file is used to load the trained model, which is then assembled using the same settings as during training.
- The predict function, which creates predictions for each input sample, receives the preprocessed test data and feeds it into the model.
- To assess the model's performance, the predictions are contrasted with the actual labels of the test data. To evaluate the model's efficiency in classifying emotions from audio samples, metrics like accuracy can be generated.

Deployment

- The graph function is also used to generate a bar chart that displays the anticipated probabilities for each emotion type.
- The anticipated emotion labels and associated probabilities are then shown, giving information about how well the model performed on previously unknown data.

VII. RESULTS AND DISCUSSIONS

Methods	Datasets	Accuracy in %
RNN	Audio Set	59.70%
HMM	Common Voice	60.71%
LPC	Audio MNIST	71.02%
MFCC	SAVEE	87.92%
Hybrid and other Methods	SAVEE	62.81%
Proposed Method (CNN)	SAVEE	89.33%

Table 1.1 Accuracy of detection for different methods and datasets

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

Table1.1 depicts the accuracy of detection(in %) for different methods and datasets

In the proposed method the accuracy is calculated by giving correct predictions a weight of 10 and erroneous predictions a weight of 1.

By adding together all of the weights for the forecasts and dividing that total by the total number of predictions, it is possible to determine the overall accuracy.

Accuracy = (Sum of Weights) / (Total Number of Predictions)

CNNs have been widely used for emotion recognition from speech, and they have achieved promising results. CNNs are capable of learning high-level representations from raw speech signals, and they can capture both spectral and temporal features that are important for emotion recognition.

Several studies have used CNNs to classify emotions into discrete categories, such as happy, sad, angry, and neutral. For example, a study by Zhang et al. (2019) used a CNN-based model to recognize emotions from speech signals. They achieved an accuracy of 73.3% on the Berlin Emotional Speech Database, which contains emotional speech samples from 10 different actors.

Overall, CNNs have shown great potential for emotion recognition from speech, and they have achieved state-of-the-art performance in many cases. However, the performance of CNNs can be affected by various factors, such as the size and quality of the training data, the choice of hyperparameters, and the preprocessing techniques used. Therefore, careful experimentation and tuning are necessary to obtain the best results.

VIII. CONCLUSION AND FUTURE WORK

In this paper, a method that uses deep learning technique for recognizing emotion from speech was proposed. The proposed work has detected the Emotion Recognition from speech as sad, disgust, happy, neutral, calm, angry, fearfull and surprised. In the proposed framework, CNN-LSTM fusion have been used to characterize emotional states of the acted speeches utterness using SAVEE dataset. An eight point scale of emotions is been used to express sentiment. Convolution Neural Networks are used to classify the features from speech samples. Mel Frequency Cepstral Coefficient and Spectogram based representations have been used to extract the features from audio signals. The proposed framework method that can truly improve the accuracy of the actual Emotion Recognition from speech dataset without additional conditions. Future work could focus on developing models such as categorising subtle subtleties, such as sarcasm or irony, or recognising and adapting to diverse accents.

REFERENCES

[1]. Gustavo Aguilar, Viktor Rozgic, Weiran Wang, and Chao Wang. 2019. Multimodal and multi-view models for emotion recognition. arXiv:1906.10198. Version 1.

[2]. Manas Jain et.al., 2020. Speech Emotion Recognition using Support Vector Machine.

[3]. Bagus Tris Atmaja and Akira Sasou 2022. Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations.

[4]. Navonil Majumder et.al. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations.

[5]. Soujanya Poria et.al., 2019. Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances.

[6]. Sahu, P., Routray, A., & Singh, S. K. (2021). Deep Emotion Recognition from Speech using Transfer Learning. IEEE Access, 9, 3322-3332.

[7]. Zhang, X., Wu, Z., Huang, Y., & King, I. (2017). Emotion Recognition in Speech using Gaussian Mixture Model and Deep Belief Network. IEEE Signal Processing Letters, 24(2), 279-283.

[8]. Wang, Y., Zhang, J., Lei, Y., & Li, S. Z. (2018). Speech Emotion Recognition Based on Joint Learning of Gaussian Mixture Model and Convolutional Neural Network. IEEE Transactions on Affective Computing, 9(3), 321-330.

[9]. Zhang, Y., Wang, J., & Zhao, Y. (2018). Emotion Recognition from Speech using Multi-task Learning with Deep Recurrent Neural Networks. IEEE Transactions on Multimedia, 20(10), 2776-2787.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.301
Volume 3, Issue 10, May 2023

[10]. Amiriparian, S., Kermani, P., & Busso, C. (2017). Emotion Recognition from Speech using Convolutional Neural Networks with Multi-Task Learning. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2227-2231).