JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

SPEECH EMOTION RECOGNITION SYSTEM USING CNN ALGORITHM

[1]Sandeep Koro, [2]P. Gayatri,
[1]MCA 2nd year, [2]Assistant Professor,
[1]Master of Computer Application,
[1]Sanketika Vidya Parishad Engineering College, Visakhapatnam, India

1.ABSTRACT

Speech Emotion Recognition (SER) is a developing research area aimed at classifying emotional states of speakers using their speech signals. CNNs have been widely applied in SER due to their ability to automatically extract relevant features from raw speech data. This paper presents a convolutional neural network model for speech emotion recognition. The proposed system consists of four main elements: emotion classification, data augmentation, model training, and feature extraction. Data augmentation involves applying various signal processing methods, such as pitch shifting, time stretching, noise addition, and dynamic range alteration, to the original speech signals. To capture the spectral information of the speech signals, the feature extraction module utilizes Mel-Frequency Spectral Coefficients (MFCCs) and their first and second-order derivatives. The proposed system is trained and evaluated using two datasets, namely RAVDESS and TESS. The emotion classification module employs a CNN architecture to categorize emotions into seven basic emotion categories: sad, fear, surprise, angry, happy, neutral, and disgust.

Keywords: Speech Emotion Recognition, Data Augmentation, MFCC, CNN, RAVDESS, TESS.

2. INTRODUCTION

Speech is a natural form of expression, particularly crucial in remote communication, where the ability to recognize and interpret emotions conveyed through speech becomes vital. However, identifying emotions in speech poses challenges due to the subjective nature and lack of consensus on quantifying and categorizing them. Speech Emotion Recognition (SER) systems utilize various techniques to process and classify speech signals, enabling the detection of embedded emotions. These systems find applications in analyzing interactions between callers and agents or in interactive voice assistants. This research aims to analyze the resonance characteristics of recorded audio to unveil hidden emotions within speech. Speech carries diverse information, such as speaker identification, age, gender, locality, and emotions. Emotions play a significant role in expressing feelings, with speech characteristics varying based on the speaker's emotional state. Although humans can instinctively recognize a speaker's emotional state, implementing this ability in machines is a complex task. Emotion detection in speech involves extracting acoustic features such as pitch, loudness, spectral characteristics, and duration from the audio signal. These features are then used to train deep learning algorithms, enabling the classification of expressed emotions. The efficacy of the deep learning algorithm, coupled with the quality and quantity of extracted acoustic features, determines the overall performance of the system.

Deep learning, as a concept, can be likened to the human nervous system. In the context of machine vision, deep learning models are utilized to learn from datasets containing images or audio, similar to how humans perceive visual information. Various deep learning models have been developed to impart a computer with visual perception capabilities similar to humans. Each node within a deep learning network acts as a neuron within a larger network, mirroring the intricacies of the human nervous system. Deep learning models constitute a subset of artificial neural networks, wherein algorithms progressively analyze input audio or images at deeper levels as

they traverse through the network layers. The prediction ultimately reaches an optimal node, generating an output that aligns with expectations.

3. LITERATURE SURVEY

Swain et al. focused on databases, feature extraction, and classifiers in their work from 2000 to 2017, neglecting neural networks and deep learning methods for SER systems. While their research comprehensively evaluated databases and feature extraction, the authors regret not exploring deep learning techniques for categorization.

In another study, Khalil et al. reviewed discrete methods in SER that utilized deep learning techniques. The study discussed the advantages and disadvantages of various deep learning approaches, including autoencoders, RNN, CNN, and DNN. However, the researchers did not address the identified limitations of these methods.

Anjali et al. recently published an overview of speech emotion detection methods. The research provides a concise examination of the different features used for recognizing speech emotions and evaluates the methods employed between 2009 and 2018. Although lacking in-depth analysis, the paper serves as a starting point for further research in this area.

Basu et al. conducted a comprehensive study in 2020, emphasizing the importance of speech emotion datasets, noise removal techniques, and various classification approaches like SVM and HMM. While the investigation identified relevant features related to SER, it lacked a thorough analysis of contemporary methods. Deep learning techniques such as CNN and RNN were briefly discussed.

Akcay et al.'s study provides comparatively in-depth analyses of records, characteristics classifiers, and emotion networks. The research evaluates machine learning methods for enhancing classifications. However, the study did not present comparative outcomes from diverse methods beyond initial findings from the respective original papers.

4. DATASETS USED

We are utilizing two datasets, namely RAVDESS and TESS, both of which can be found on Kaggle.com.

4.1. RAVDESS DATASET

RAVDESS is an acronym for the Ryerson Audio-Visual Database of Emotional Speech and Song. It consists of a total of 1440 files, obtained by having 24 actors participate in 60 trials each (60 trials per actor x 24 actors = 1440). The database features 24 professional actors, with an equal distribution of 12 female and 12 male individuals, who vocalize two statements that are lexically matched and delivered in a neutral North American accent. The speech emotions

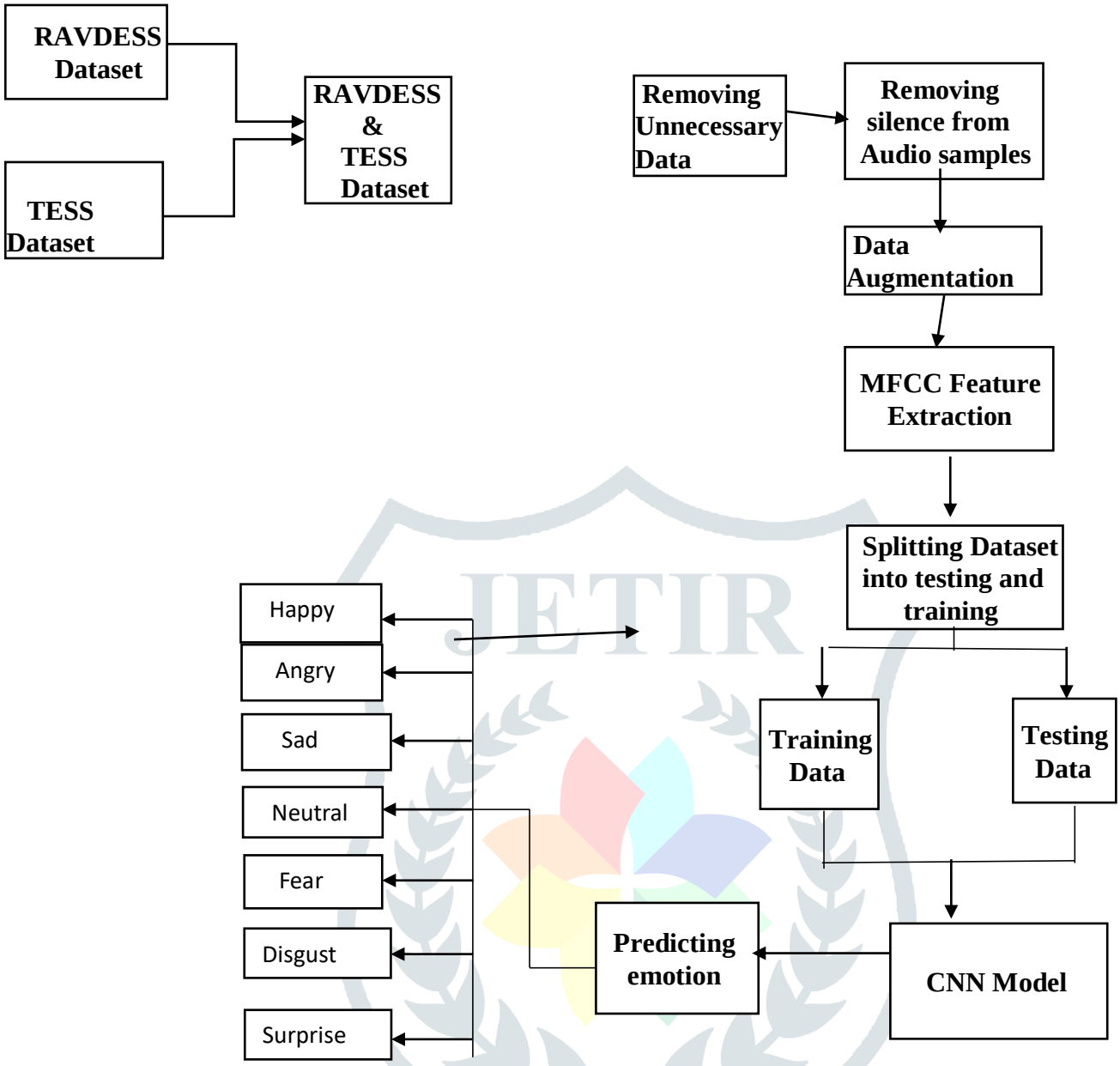
expressed in this database encompass a range of expressions, including neutral, calm, happy, sad, angry, fearful, surprise, and disgust.

4.2. TESS DATASET

A collection of 200 specific words was spoken in the introductory phrase "Say the word _" by two actresses, aged 26 and 64 years. Recordings were captured for each word, depicting seven distinct emotions: anger, disgust, fear, happiness, pleasant surprise, sadness, and neutrality. The dataset comprises a total of 2814 data points, represented by audio files.

The dataset is structured in a manner where each individual actress and their corresponding emotions are organized within separate folders. Within each folder, all 200 audio files of the target words can be found. These audio files adhere to the WAV format.

5. SYSTEM ARCHITECTURE



6. METHODOLOGY

6.1 Data collection:

The initial step in implementing the Speech Emotion Recognition system involves gathering audio samples categorized under different emotions, which will be used for model training. For this system, the audio samples are sourced from the RAVDESS and TESS datasets.

6.2 Elimination of unnecessary audio and silence:

Unwanted audio samples are removed from the dataset, leaving only the audio samples corresponding to emotions such as happiness, anger, sadness, neutrality, fear, disgust, and surprise. Additionally, silence is eliminated from each audio sample to reduce unnecessary data. The Librosa.effects.trim function is applied to remove silence from the audio samples, with any sounds below 30 decibels being removed.

6.3 Data Augmentation:

Data augmentation is a widely used technique in deep learning that expands the training data size by generating new data based on existing data. In speech emotion recognition, data augmentation enhances the performance of deep learning models by increasing the quantity and diversity of training data.

6.4 Feature Extraction:

Feature extraction involves transforming raw data into numerical features that can be processed while retaining the essential information from the original dataset. It yields superior results compared to directly applying algorithms to raw data. In this system, Mel-Frequency Cepstral Coefficients (MFCC) are utilized for feature extraction. The MFCC technique involves windowing the signal, applying the Discrete Fourier Transform (DFT), taking the logarithm of the magnitude, warping the frequencies on a Mel scale, and finally applying the inverse Discrete Cosine Transform (DCT).

6.5 Model Training:

The subsequent step is to train the Convolutional Neural Network (CNN) model using the data collected through MFCC feature extraction.

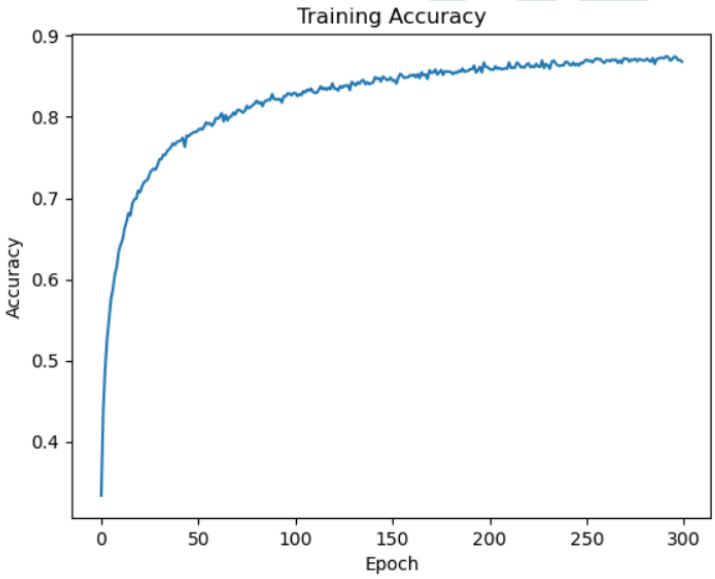
6.6 Classification of speech emotions:

When an audio file is imported, it is passed through the CNN model to detect the emotion conveyed by the audio.

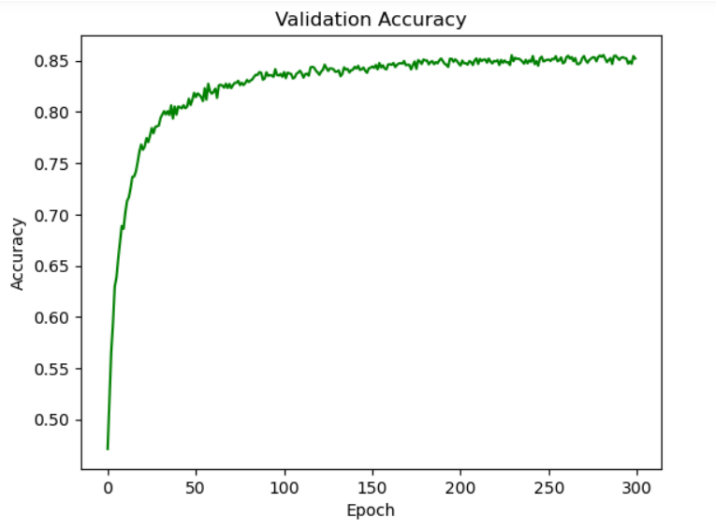
7.RESULTS

Here is an overview of the model we propose.

```
model.summary()
Model: "sequential_1"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d_2 (Conv2D)           (None, 229, 3, 64)        640

 conv2d_3 (Conv2D)           (None, 229, 3, 64)        36928

 max_pooling2d_1 (MaxPooling  (None, 115, 2, 64)       0
 2D)

 flatten_1 (Flatten)         (None, 14720)             0

 dropout_2 (Dropout)         (None, 14720)             0

 dense_2 (Dense)             (None, 128)               1884288

 dropout_3 (Dropout)         (None, 128)               0

 dense_3 (Dense)             (None, 7)                 903

=================================================================
Total params: 1,922,759
Trainable params: 1,922,759
Non-trainable params: 0
_________________________________________________________________
```



Fig, Training accuracy



Fig, Validation accuracy

8. CONCLUSION

Within this project, we have constructed a Speech Emotion Recognition (SER) system using a Convolutional Neural Network (CNN) to categorize an individual's emotional state based on their audio data. The CNN model was specifically developed to automatically extract relevant features from the original audio data and learn to predict emotional categories through supervised learning. Through our experiments, we observed that the proposed SER system exhibited a high level of effectiveness, achieving an impressive accuracy rate of 85% on the testing dataset. The CNN model surpassed traditional machine learning methods commonly employed in the field, underscoring its efficacy in SER tasks. It is worth noting, however, that the model's performance could potentially have been further enhanced with a larger and more diverse dataset.

9. REFERENCES

1. Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.

2. Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. https://www.mdpi.com/1424-8220/21/4/1249/htm.

3. Vocal Technologies, 'Pitch Detection using Cepstral Method', last accessed 25th February 2019 url: https://www.vocal.com/perceptual-filtering/pitch-detection/.

4. A Study on the Search of the Most Discriminative Speech Features in the Speaker Dependent Speech Emotion Recognition. https://dl.acm.org/doi/10.1109/PAAP.2012.31.

On the Speech Properties and Feature Extraction Methods in Speech Emotion Recognition. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7962835/.

What is the difference between categorical, ordinal and interval variables?. Last accessed 27thFebruary2019Url:https://stats.idre.ucla.edu/other/mult-pkg/whatstat/what-is-the-difference-between-categorical-ordinal-and-interval-variables/

Sandeep Koro is studying his 2nd year, Master of Computer Applications in Sanketika Vidya Parishad Engineering College, affiliated to Andhra University, accredited by NAAC. With his interest in python and CNN model and as a part of academic project, he chooses Speech Emotion Recognition system using CNN model. As a result of our analysis, we discovered interesting statistics that can help to predict a emotion of a person. A completely developed project along with code has been submitted for Andhra University as an Academic Project. In completion of his MCA.

Potnuri Gayatri: Assistant professor, she received her M Tech in Computer Science & engineering from JNTU Kakinada in January 2015. She received her B Tech Degree from VITAM College of Engineering, JNTU Hyderabad in 2004. she currently working as Assistant Professor, CSE Dept in SVPEC, Andhra Pradesh, India. Her Research Interests include Sensor Networks.