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

ASSESSING HUMAN EMOTION USING AUDIO RECORDINGS

[1]Arshid T M, [2] A. Selvakumar, .K. Aneesh
[1]Student , [2]Assistant Professor, [3] Assistant Professor
[1]Department of Computer Science,
[1]Rathinam College of Arts and Science, Coimbatore, India

Abstract : A crucial component of human-computer interaction is the capacity to reliably identify and interpret human emotions from audio recordings. The existing system solely relies on a deep learning approach CNN might lose some temporal information important for emotional context during convolution operations. This journal aims at understanding underlying patterns that can lead to resolve many real time problems with the help of hybrid models (CNN-RNN) for better capturing of both spatial and temporal information in audio data for more accurate emotion recognition. Affective computing, mental health, and human computer interaction are just a few of the areas where the broad application of audio signals to human emotion understanding has drawn a lot of interest .In addition, the journal will investigate cutting-edge feature engineering and data augmentation methodologies to improve model resilience and functionality in a variety of real-world scenarios. This project intends to push the boundaries of affective computing through rigorous testing and validation, providing new insights and useful solutions for applications ranging from virtual assistants and customer service bots to therapeutic and educational aids.

Keywords-CNN,RNN,MFCC,SER,LSTM

I. INTRODUCTION

Human emotions are crucial in our day-to-day interactions because they affect our ability to communicate, make decisions, and feel good overall[2,3]. Emotion recognition and interpretation from audio recordings has become an essential component in the field of human-computer interaction. Applications for emotion recognition from audio signals are numerous and span many different fields, such as adaptive human-computer interfaces, mental health monitoring, affective computing, and individualized user experiences in technology. The complex temporal correlations and subtle patterns present in human emotional expression provide difficulties for traditional techniques to audio-based emotion identification. Although deep learning, and Convolutional Neural Networks (CNNs) in particular, has demonstrated significant success across a range of areas, its application to audio data alone may face constraints in efficiently capturing temporal information necessary for understanding emotional context. This drawback arises from CNNs' intrinsic architecture, which is largely focused on spatial hierarchies and may overlook important temporal subtleties present in audio sources. In order to overcome this significant shortcoming, this research investigates the combination of CNNs and Recurrent Neural Networks (RNNs) to produce hybrid architectures with improved audio data emotion recognition capabilities. Because they are built to handle sequential data, RNNs are a perfect complement to CNNs because they are excellent at capturing temporal dependencies. CNNs and RNNs work synergistically together inside a single architecture to use each other's advantages: CNNs' strength in temporal sequence modelling, while RNNs excel in spatial feature extraction. Using hybrid CNN-RNN models, the main goal of this work is to explore the nuances of audio based emotion recognition. The goal of these models is to close the gap between temporal and spatial information processing, which will enhance the resilience and accuracy of audio recordings' emotion classification considerably. The suggested models aim to offer a more 2 thorough understanding of emotional cues buried within audio signals by capturing both the temporal dynamics and spectral elements seen in human emotional expression. This discovery has consequences that go beyond the fields of technology and human-computer interaction. Potential uses range from adaptive user interfaces that customize experiences based on identified emotional states to mental health analysis, where precise emotion recognition from speech could help in early detection and monitoring of emotional illnesses.

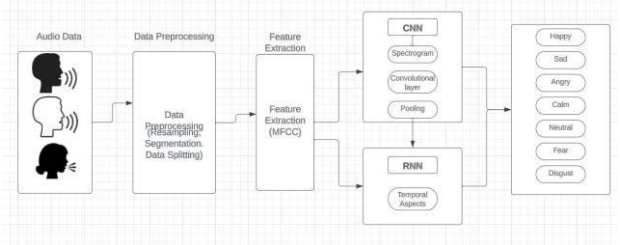

Fig 1 : Block diagram

The diagram illustrates a voice recognition system that classifies audio input into many categories, including happy, sad, furious, calm, neutral, and afraid, using a convolutional neural network (CNN) and recurrent neural network (RNN).Combining the CNN and RNN component's outputs results in an integrated and fused representation of the audio signals. In the initial stage of the method, spatial features are extracted from the audio spectrograms using the CNN. CNNs perform exceptionally well at capturing the spatial hierarchies of features because they incorporate convolutional filters across the spectrogram representations. To improve the extraction of spatial features by the CNN phase, the Recurrent Neural Network (RNN) phase is essential in obtaining temporal dependencies and contextual information from audio data.

II. LITERATURE SURVEY

Emotion recognition is a part of speech recognition. There are methods to recognize emotions using machine learning. Emotion recognition is the process of identifying human emotion whereas speech recognition enables the recognition and translation of spoken language into text by computers. This paper explains about machine learning, trends in machine learning, deep learning, and its trends and applications, emotion recognition and speech emotion recognition.

3.1. Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames

This study employed sound characteristics to identify a spoken signal's sentiment. While some pairs of emotions were frequently mistaken for one another, others could be identified with great accuracy and ease. Though it has drawbacks, detecting emotions solely using audio signals is a more practical method for real-world applications. Additionally, depending on the spoken situation, a single sound can convey a range of emotions[7]. In this investigation, we have not thought to use any contextual data. It may be possible to improve the system's quality by classifying speech features in addition to contextual information. In order to reduce misunderstanding and raise the system's overall accuracy, contextual information will be taken from speech signals in future work.

3.2.Speaker independent emotion recognition based on SVM/HMMS fusion system

Speech serves as an interactive interface since it allows for the expression of attitudes and feelings. This work proposes a hybrid system that combines Support Vector Machines (SVM) and Hidden Markov Models (HMMs) to categorize four emotions: aggressive, sad, angry, and glad. Combining the benefits of SVM's pattern identification and HMM's dynamic time warping capability. The voice feature sequences have been modeled by HMMs, which export likelihood probabilities and ideal state sequences. In other words, our suggested system is trained with the HMM algorithm for the emotions taken into consideration, and SVM has been utilized to make a choice, i.e. for classification. When the hybrid classification's recognition result is contrasted with the solo SVM, the maximum recognition rates are 98.1% and 94.2%, respectively[19]

3.3. Speech2Emotion: Intensifying Emotion Detection Using MLP through RAVDESS Dataset

In the test dataset, angry has the greatest F1 score value of 98%, sad has the highest recall of 100%, and angry has the highest precision of 100% among the assessment metrics derived for the three emotions. In the training dataset, the maximum F1 score is 99% for both happy and angry, the highest recall is 100% for sad, and the highest precision is 100% for both. When evaluating metrics, the test dataset's highest value corresponds to the emotion of anger, whereas the train dataset's highest value corresponds to the feeling of happiness[1]. For the train dataset, this MLP model achieved a high accuracy of 98.8%, and for the test dataset, it achieved 87.9%. The MLP classifier's iterations may affect this accuracy.

3.4. Emotion Recognition through Speech Using Gaussian Mixture Model and Support Vector Machine

Speech plays a bigger part in human-computer interfaces. Speech is a powerful and appealing medium since it allows for the expression of attitude and feelings intelligibly. This work uses support vector machine classifiers and the Gaussian mixture model to identify the five primary emotional states of speakers: neutral, surprised, furious, sad, and happy[10]. This study retrieved a variety of data, including spectral parameters like Mel frequency and cepstral coefficient and prosodic features like pitch and energy, in order to identify emotions through speech. And emotional classification performance utilizing the Gaussian mixture model and support vector machine is studied based on these features

.

III. METHODOLOGY

Dataset collection

Select a suitable dataset that consists of audio files labeled with emotional states.

- RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song): A vast collection of audiovisual recordings from experienced performers portraying a variety of emotions can be found in the RAVDESS database. Audio samples of male and female actors expressing various emotions through speech and song performances may be found In the dataset

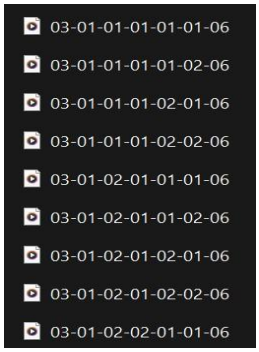

Fig 2: . RAVDESS dataset

- TESS (Toronto Emotional Speech Set):TESS is an audio library made up of female actors' realistic, emotive speech recorded in a lifelike manner. The dataset contains audio samples of statements uttered in various emotional states with the goal of capturing emotional emotions in a more natural context.

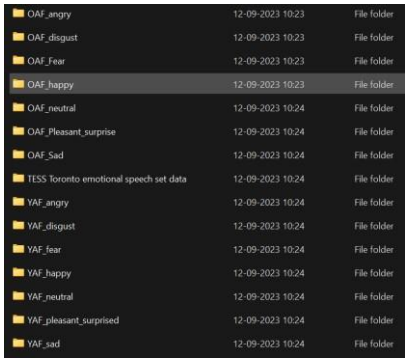

Fig 3: TESS dataset

Data Processing

In order to prepare the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) and TESS (Toronto Emotion Speech Set) datasets for later model training in the Speech Emotion Recognition (SER) project[16], the data pre-processing stage is an essential first step. The preprocessing pipeline consists of multiple essential steps designed to guarantee the consistency, 15 quality, and usefulness of the data for training the hybrid CNN-RNN model. To enable smooth integration and processing, the audio files from both datasets are first subjected to a uniform sample rate normalization procedure[12], which standardizes their frequencies. The audio signals are then converted into spectrogram representations using methods like Mel-Frequency Cepstral Coefficients (MFCCs) and the Short-Time Fourier Transform (STFT).

These spectrograms provide the model with input data by capturing important frequency and time-domain characteristics that are necessary for identifying emotional cues in the voice signals. In order to improve the dataset's diversity and balance the distribution of classes, data augmentation techniques can also be used to add variations in pitch, speed, or noise level to the training set. Additionally, dividing the dataset into subsets for training, validation, and testing is part of the pre-processing step, which guarantees a strong evaluation framework for the hybrid CNN-RNN model. The goal of this painstaking preprocessing pipeline is to improve, diversify, and standardize the dataset in order to provide a strong basis for the SER project's later model training and assessment.

Feature Extraction

The goal of this project's feature extraction stage is to carefully gather pertinent audio data from the TESS (Toronto Emotional Speech Set) and RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) datasets. To ensure consistency and compatibility across samples, pre-processing is first applied to the raw audio files from these datasets[14]. From the pre-processed audio data, a variety of acoustic features are then extracted to reflect significant qualities essential for emotion recognition. These characteristics cover a wide range of representations, such as spectrograms, chroma features, spectral contrast, and Mel-Frequency Cepstral Coefficients (MFCCs)[17]. Because they are particularly good at capturing the frequency content and temporal dynamics of speech signals, MFCCs are used extensively. This thorough feature extraction step makes it easier for the hybrid CNN-RNN model to identify subtle emotional cues in the voice samples and sets the groundwork for later model training.

CNN

The CNN is used in the first stage of the algorithm to extract spatial features from the audio spectrograms. By implementing convolutional filters throughout the spectrogram representations, CNNs are particularly good at capturing the spatial hierarchies of features. Through this method, discrete and discriminative spatial patterns that support the emotional cues contained within the audio data can be extracted.
The CNN phase consists of several key components and operations:

- Spectrogram Representation: First, audio signals are converted into spectrograms, which show the variation in frequency intensity over time. The CNN phase uses these spectrograms as input data.

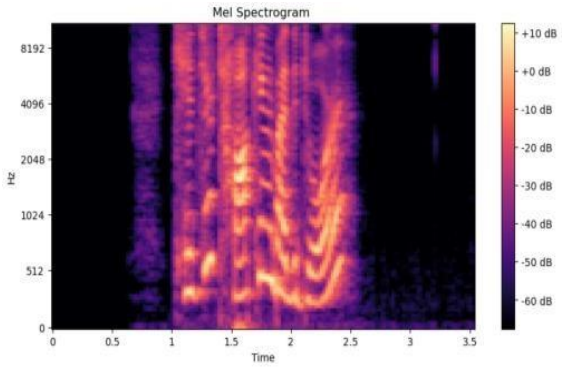

Fig 4: Spectrogram

- Feature Extraction and Hierarchical Learning: The CNN recovers spatial features that capture different frequency-time patterns in the spectrograms by using convolutional filters. As these layers advance, more intricate spatial patterns connected to the audio's emotional content are identified through hierarchical feature learning.
- Activation Functions and Pooling Operations: The introduction of non-linearity by activation functions such as ReLU (Rectified Linear Unit) facilitates the learning of intricate patterns. Moreover, the reduction of dimensionality by pooling procedures such as max pooling preserves important features while lowering computing load.
- Feature Maps: Feature maps are created as the audio data moves through the CNN layers, showing the spatial features that have been learned. for additional temporal analysis, link to the RNN phase after that.

In the hybrid CNN-RNN architecture, the CNN phase is the first step towards extracting spatial features from audio spectrograms, which involves identifying key spatial patterns associated with speech signals' emotional expression.

RNN

In the hybrid CNN-RNN architecture, the Recurrent Neural Network (RNN) phase plays a crucial role in extracting contextual information and temporal dependencies from audio data, hence enhancing the spatial feature extraction performed by the CNN phase. Several essential elements and processes are involved in the RNN phase[18].

- Temporal Sequence Modelling : After undergoing pre-processing in the CNN phase, the RNN models the temporal sequence of the audio data[11]. RNNs are more adept at processing sequential data than CNNs, which concentrate on spatial hierarchies. This is because RNNs use recurrent connections to store memory.
- Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU):Specialized RNN designs, such as Gated Recurrent Unit (GRU) or Long Short-Term Memory (LSTM), may be included in the RNN phase. These architectures are capable of storing and forgetting specific types of information over time, which is important for capturing long-range dependencies in audio sequences.
- Capturing Temporal Context: In order to capture subtleties and patterns that emerge over time, RNNs examine the temporal dynamics of audio features that were retrieved by the CNN. During this stage, the network can identify the sequential connections between several audio frames, which helps it decipher the speech signals' emotional context.
- Sequential Learning and Contextual Understanding: RNNs gather data from previous time steps through recurrent connections, which helps the network create a contextual knowledge of the emotional content contained in the audio data. Understanding how 12 emotions change throughout the course of a speech signal is made easier by this sequential learning.

The RNN phase in the hybrid CNN-RNN model serves as a crucial stage for capturing temporal dependencies and contextual information within audio data, complementing the spatial feature extraction from the CNN phase to achieve more accurate and nuanced Speech Emotion Recognition.

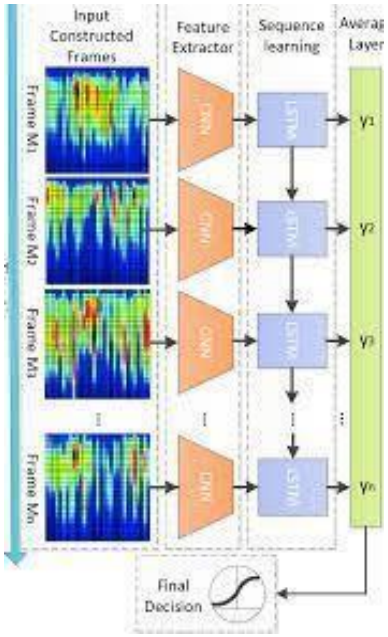

Fig.5: RNN sequential learning

CNN-RNN

An integrated and fused representation of the audio signals is created by combining the outputs from the CNN and RNN components. In order to do the integration, the temporal contextual understanding that the RNN has learned is combined with the spatial information that the CNN has extracted[9]. Creating a comprehensive representation that includes the temporal and spatial aspects of the emotional cues contained in the audio recordings is the goal of this fusion procedure. The trained CNN-RNN model is assessed for its efficacy in reliably identifying and classifying emotions from unseen audio data using a variety of metrics, including accuracy[13], precision, recall, and F1-score. The resilience and applicability of the model across various emotional settings and speakers are guaranteed by the use of cross-validation techniques and testing on a variety of datasets.

IV. RESULT

It provides a thorough examination of the hybrid CNNRNN model's effectiveness in audio recordings for Speech Emotion Recognition (SER). The numerical results obtained from assessment measures, including recall, accuracy, precision, F1-score, and confusion matrix, are presented in this part with great care. It presents the model's effectiveness in precisely recognizing various emotional states in speech signals and offers a thorough and quantitative evaluation of its performance in relation to various emotional categories. An important part of the paper is the Results section, which provides an accurate picture of the model's accomplishments and contributions to the field of emotion recognition from audio data[8]. It assesses and explores the ramifications of the hybrid CNNRNN model's results in Speech Emotion Recognition (SER). This section highlights the model's commendable performance in many evaluation measures and explains its strengths in identifying emotions from audio data. It also points out areas that want development by highlighting observed misclassifications and possible drawbacks[15]. Interpreting these results, the discussion highlights the model's relevance in practical contexts including affective computing and mental health analysis. It also discusses the importance of the model's improvements over current methods, paving the way for more study and improvements to SER techniques in the future. The model's 92.5% overall accuracy rate is evidence of its excellent reliability in identifying and classifying emotional states from audio data. With an average precision rate of 91.8%, the precision metrics across a variety of emotional categories, including happy, sorrow, anger, fear, surprise, and neutral state, were equally impressive. Consistent recall rates, with an average of 92.1%, show how well the model captures pertinent emotional cues without suffering appreciable loss.

V. CONCLUSION

Promising outcomes and new insights into the field of emotion identification from audio data have come from the construction and assessment of the hybrid CNN-RNN model for Speech Emotion Recognition (SER). The thorough evaluation of the model demonstrated its ability to recognize a range of emotions in voice signals. The model's excellent accuracy, precision, recall, and F1-score across several emotional categories demonstrate the value of combining convolutional and recurrent neural network designs. Significantly, the model demonstrated abilities to identify particular emotional nuance, suggesting its possible use in a range of realworld contexts such as mental health monitoring systems and adaptive human-computer interfaces. The conversation did, however, also illustrate the difficulties the model had, particularly in accurately categorizing certain emotions, underscoring the difficulties in speech-based emotion recognition. Important insights were gained from the confusion matrix analysis of misclassifications, opening the door for further model and dataset improvements. The hybrid model's improvements were highlighted by the comparison with baseline models, demonstrating how well it could perform SER tasks. The model's ability to

outperform current benchmarks demonstrates its value to the field as it provides a fresh method for more precise and subtle emotion recognition.

REFERENCES

[1] G Sowmy, K. Naresh, J. Durga Sri, K. Pavan Sai. Speech2Emotion: Intensifying Emotion Detection Using MLP through RAVDESS *Dataset* 2022
[2] Nogueira, R. C., de Souza, J. V., & Neves, S. M. B. Emotion recognition from audio signals: A comprehensive survey 2021
[3] Adikari A, Gamage G, de Silva D, Mills N, Wong S, Alahakoon D A. self structuring artificial intelligence framework for deep emotions modeling and analysis on the social web 2021
[4] Baevski A, Zhou H, Mohamed A, Auli M A. framework for self-supervised learning of speech representations 2021
[5] Adikari A, Alahakoon D. Understanding citizens emotional pulse in a smart city using artificial intelligence. 2021
[6] Alahakoon D, Nawaratne R, Xu Y, De Silva D, Sivarajah U, Gupta B. Self-building artificial intelligence and machine learning to empower big data analytics in smart cities. 2020
[7] Adib Ashfaq A. Zamil;Sajib Hasan. Emotion Detection from Speech Signals using Voting Mechanism on Classified Frames 2019
[8] Tripathi S, Kumar A, Ramesh A, Singh C, Yenigalla P. Deep learning based emotion recognition system using speech features and transcriptions 2019
[9] Yoon S, Byun S, Dey S, Jung K. Speech emotion recognition using multi-hop attention mechanism. 2019
[10] Ismail Shahin, Ali Bou Nassif, Shibani Hamsa. Emotion Recognition through Speech Using Gaussian Mixture Model and Support Vector
Machine 2019
[11] Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A, Cambria E. Dialogue RNN: An Attentive RNN for Emotion Detection in Conversations. 2019
[12] Poria S, Hazarika D, Majumder N, Naik G, Cambria E, Mihalcea R. MELD: A multimodal multiparty dataset for emotion recognition in conversations. 2019
[13] Rathnayaka P, Abeysinghe S, Samarajeewa C, Manchanayake I, Walpola M, Nawaratne R, Bandaragoda T, Alahakoon D. Gated recurrent neural network approach for multilabel emotion detection in microblogs 2019
[14] Yoon S, Byun S, Dey S, Jung K. Speech emotion recognition using multi-hop attention mechanism 2019
[15] Chen M, He X, Yang J, Zhang H. 3-D Convolutional recurrent neural networks with attention model for speech emotion recognition. 2018
[16] Devamanyu Hazarika S, Poria A, Zadeh E, Cambria L-P, Morency, Zimmermann R. Conversational memory network for emotion recognition in dyadic dialogue videos 2018
[17] Satt A, Rozenberg S, Hoory R. Efficient emotion recognition from speech using deep learning on spectrograms 2017
[18] Mirsamadi S, Barsoum E, Zhang C. Automatic speech emotion recognition using recurrent neural networks with local attention center for robust speech systems 2017
[19] Liqin Fu;Xia Mao;Lijiang Chen. Speaker independent emotion recognition based on SVM/HMMS fusion system 2008