International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 10
DOI: https://doi.org/10.17762/ijritcc.v11i10.8472
***Article Received:** 26 July 2023 **Revised:** 12 September 2023 **Accepted:** 01 October 2023*

Deep Learning-Based Speech Emotion Recognition Using Librosa

D. Lakshmi[1], R.Vijay[2], R. Thalapathi Rajasekaran[3], A. Vani Lavanya[4], R. Bhavani[5]
[1]Department of Computer Science and Engineering,
Panimalar Engineering college, Chennai, India-600123
dlakshmicsepit@@gmail.com
[2]Department of Computer Science and Engineering,
Vel Tech Rangarajan Dr.Sagunthala R&D Institute of Science and Technology, Chennai, India- 600054
drvijayr@veltech.edu.in
[3]Department of Computer Science and Engineering,
Saveetha Schools of Engineering, Saveetha Institute of Medical And Technical Sciences,
Thandalam, Chennai, India- 602105
r.rajthalapathi@gmail.com
[4]Department of Computer Science and Engineering,
St.Joseph's Institute of Technology, Chennai, India- 600119
vanilavanya8@gmail.com
[5]Department of Computer Science and Engineering,
Chennai Institute of Technology, Chennai, India- 600069
bhavanir@citchennai.net

Abstract— Speech Emotion Recognition is a challenge of computational paralinguistic and speech processing that tries to identify and classify the emotions expressed in spoken language. The objective is to infer from a speaker's speech patterns, such as prosody, pitch, and rhythm, their emotional state, such as happiness, rage, sadness, or frustration. In the modern world, one of the most crucial marketing tactics is emotion detection. For a person, you might tailor several things in order to best fit their interests. Due to this, we made the decision to work on a project where we could identify a person's emotions based just on their speech, allowing us to handle a variety of AI-related applications. Examples include the ability of call centers to play music during tense exchanges. Another example might be a smart automobile that slows down when someone is scared or furious. In Python, we processed and extracted features from the audio files using the Librosa module. A Python library for audio and music analysis is called Librosa. It offers the fundamental components required to develop systems for retrieving music-related information. Because of this, there is a lot of potential for this kind of application in the market that would help businesses and ensure customer safety.

Keywords- Speech Emotion Recognition, Computational Paralinguistic, Emotion Categorization, Prosody, Audio File Processing, Voice-based Emotion Detection.

I. INTRODUCTION

Given that speaking is one of the most natural ways for humans to express themselves. We rely on it so much that, while using other channels of communication, such as emails and texts, where we frequently utilize emojis to convey the feelings we are feeling, we realize how important it is. Since emotions are a major component of communication, it is critical to identify and analyze them in the digital world of distant communication that we live in today. Since emotions are irrational, detecting them might be difficult. There's no universal agreement on how to quantify or classify them. A SER system is defined as an assembly of techniques that analyze and categorize speech signals in order to identify the emotions that are present in them. Numerous application domains, such as interactive voice assistants and caller-agent interaction analysis, can benefit from the implementation of such a system. In this work, we analyze the acoustic characteristics of the audio data of recordings in an effort to identify the underlying emotions in voice recordings.

The goal of speech processing and computational paralinguistic is to identify and classify the emotions conveyed in spoken language. This process is known as speech emotion recognition. The objective is to infer from a speaker's speech patterns, such as prosody, pitch, and rhythm, their emotional state, such as happiness, rage, sadness, or frustration. A speech's lexical characteristics (the vocabulary used), visual features (the speaker's expressions), and acoustic features (sound qualities like pitch, tone, jitter, etc.) are the three categories of features. Analyzing one or more of these characteristics can help solve the speech emotion recognition challenge. If one wishes to predict emotions using real-time audio, they would need to extract text from the voice, which would require a transcript if they choose to follow the lexical characteristics. Comparably, analyzing visual features would require additional video of the conversations, which might not always be possible. However, analyzing acoustic features would only require audio data, allowing us to analyze them in real-time while the conversation is happening.

The goal of this research is to identify and classify the emotions communicated in spoken language using speech processing and computational paralinguistic. The objective is to infer from a speaker's speech patterns, such as prosody, pitch, and rhythm, their emotional state, such as happiness, rage, sadness, or frustration. We will be able to handle several AI-related applications when we have this data. Examples include the ability of call centers to play music during tense exchanges. Another example might be a smart automobile that slows down when someone is scared or furious. Because of this, there is a lot of potential for this kind of application in the globe, which would help businesses and give customers protection.

II. RELATED WORKS:

The "Concurrent Spatial-Temporal and Grammatical (CoSTGA)" model, a deep learning architecture intended to concurrently capture spatial, temporal, and semantic representations, is introduced by the authors in this study. Using a two-level feature fusion strategy, this model combines related features from several modalities at the local feature learning block (LFLB) in the first level. They also provide the "Multi-Level Transformer Encoder Model (MLTED)" for contrasting single-level and multi-level feature fusion. Through its multi-level approach, the CoSTGA model demonstrates better model efficacy and resilience by efficiently integrating spatial-temporal characteristics with semantic trends [1]. This research uses both single-task and multitask learning techniques to evaluate speech emotion and naturalness recognition using deep learning models. The emotion model takes dominance, valence, and arousal into account, and multitask learning predicts naturalness scores at the same time. When it comes to forecasting extreme scores, the model is limited. However, when it comes to jointly predicting naturalness, future emotion recognition algorithms may do better [2]. The "autoencoder with emotion embedding," a novel technique for extracting deep emotion characteristics from voice data, is presented in this study. This model uses instance normalization and makes use of emotion embedding, in contrast to other efforts that used batch normalization, to help the model learn emotion-related data effectively. Through data augmentation, the method improves generalization by fusing acoustic characteristics from the openSMILE toolbox with latent representations from the autoencoder. Comparing IEMOCAP and EMODB evaluation results to other spoken emotion recognition systems, significant performance gains are seen [3].

In order to improve Speech Emotion Recognition (SER), this work combines self-attention and bidirectional long short-term memory (BLSTM) techniques. It presents a brand-new strategy with confidence metrics called Self-Attention Weight Correction (SAWC). In order to lessen the impact of speech recognition problems in text features and to highlight segments with similar faults in acoustic features, SAWC is used to both the text and acoustic feature extractors in SER. This technique outperforms previous state-of-the-art algorithms with a weighted average accuracy of 76.6%, as demonstrated by experiments on the IEMOCAP dataset. Its performance is explored in detail across feature extractors [4]. In order to improve Speech Emotion Recognition (SER), this work combines self-attention and bidirectional long short-term memory (BLSTM) techniques. It presents a brand-new strategy with confidence metrics called Self-Attention Weight Correction (SAWC). In order to lessen the impact of speech recognition problems in text features and to highlight segments with similar faults in acoustic features, SAWC is used to both the text and acoustic feature extractors in SER. This technique outperforms previous state-of-the-art algorithms with a weighted average accuracy of 76.6%, as demonstrated by experiments on the IEMOCAP dataset. Its performance is explored in detail across feature extractors [5]. This research presents a novel method for Speech Emotion Recognition (SER) based on a deep neural network (DNN) with an attentive temporal pooling module. This module automatically highlights the emotionally charged sections and downplays the less important ones. The Gaussian Mixture Model (GMM) and an additional DNN are used to extract emotional saliency weights from condensed representations. Surprisingly, our methodology relies just on utterance-level labels to achieve state-of-the-art SER performance on many public emotion datasets, including as RML, EMO-DB, and IEMOCAP, without requiring supervisory information at the frame or segment level [6].

This research provides a hybrid technique for speech emotion recognition (SER) that combines Transformer Encoder and Long Short-Term Memory (LSTM) Network. It provides notable gains in recognition by utilizing aspects of the Mel Frequency Cepstral Coefficient (MFCC). On the RAVDESS, Emo-DB, and language-independent datasets, the hybrid LSTM-Transformer model achieves recognition success rates of 75.62%, 85.55%, and 72.49%, outperforming previous models in the field. This novel method improves speech signal long-term dependency learning and advances emotion categorization [7]. Speech data research has become more important during the last ten years, particularly in the areas of entertainment, security, healthcare, and human-computer interaction. The TLEFuzzyNet model, a three-stage pipeline for speech emotion recognition, is presented in this research. In the first step, it uses data augmentation and Mel spectrogram extraction; in the second, it uses pretrained CNN models for transfer learning; and in the third stage, it combines the prediction scores of the models using Fuzzy Ranks and a modified Gompertz function. Tested using RAVDESS and EmoDB datasets, the TLEFuzzyNet model delivers state-of-the-art performance, proving to be a strong foundation for Speech Emotion Recognition (SER) [8]. In the context of the Internet of Things, this study presents EdgeRNN, a small voice recognition network intended for edge computing (IoT). Recurrent neural networks (RNN) are used in EdgeRNN to process temporal information, while 1-Dimensional Convolutional Neural Networks (1-D CNN) are used to analyze spatial information. It also has a streamlined attention system. Using the IEMOCAP dataset, EdgeRNN's performance is showcased for speech emotion identification, with an unweighted average recall (UAR) of 63.98%. Furthermore, with a weighted average recall (WAR) of 96.82%, it outperforms similar efficient networks on the Raspberry Pi 3B+ in voice keyword recognition [9].

Speech Emotion Recognition (SER) has gained significant importance in sophisticated speech processing and human-computer interaction during the last ten years. Though there are major differences between human and machine techniques, SER uses feature extraction and classification to detect emotional cues in speech. This study provides insights into this developing field of research by surveying recent SER literature and connecting multidisciplinary expertise from SER, applied psychology, and HCI. By highlighting research gaps, it promotes a cutting-edge understanding of this intricate subject for academics, institutions, and regulatory agencies to take into account [10]. Within Industry 4.0, increasing operator productivity—which is heavily impacted by emotions—is essential to improving industrial processes. While negative emotions, such as irritation, have a negative correlation with productivity but a positive correlation with workplace misbehavior, good emotions, such as happiness, have a positive correlation with productivity. Using the first two convolutional layers of Baidu's 2015 speech recognition model as the foundation, this work uses a speech-based emotion identification system. Even in the absence of contextual information, the model achieves state-of-the-art performance with a 23% F1-score rate for the feelings of surprise and delight, accurately predicting seven key emotions on the MELD test set [11]. Joseph Weizenbaum created ELIZA, a groundbreaking tool for natural language processing, between 1964 and 1966. ELIZA used simulated dialogues to entice users into thinking they understood each other, particularly with its well-known "DOCTOR" script, which used non-directional inquiries to mimic person-centered treatment. Though it lacked true comprehension, many users were taken aback by ELIZA's seeming empathy, underscoring disconnect between human emotions and machine interactions in the early stages of AI research [12].

In human-computer interaction, Speech Emotion Recognition (SER) is becoming a more significant problem. Based on databases containing a single emotion label for each speech, traditional SER models provide a single emotion label for each utterance. A new emotional speech database was developed to address the complexity of numerous emotions occurring at different times and in different intensities in human speech. Of the 2,025 examples in this collection, 1,525 exhibit several emotions, which helps to create SERs that are more realistic and subtle [13]. This study explores speech emotion recognition using a combination of acoustic and linguistic features. While prior studies often relied on reference transcripts for emotional speech recognition, this research aims to extract linguistic features from speech recognition results. Despite a word recognition accuracy of 82.2% and some errors, the combination of linguistic and acoustic features proves effective for high-performance emotion recognition. This approach offers a promising way to enhance emotional speech recognition models [14]. Due to a lack of emotion-labeled data, this work tackles the problem of acquiring trustworthy linguistic characteristics for Speech Emotion Recognition (SER). For collaborative training, it suggests merging Automatic Speech Recognition (ASR) outputs into the SER pipeline. Through studies, it shows that SER performance may be greatly improved with a hierarchical co-attention fusion technique that incorporates both ASR hidden and text outputs, attaining 63.4% weighted accuracy on the IEMOCAP corpus. Furthermore, using layer-difference analysis of the Wav2vec 2.0 model and word error rate analysis, the article provides insightful information on the ASR-SER relationship [15].

III. EXISTING SYSTEM:

Text transcriptions and audio signals are analyzed in order to identify and categorize the emotional states expressed in speech. To infer the emotions in speech, it is necessary to analyze the intricate interactions between the retrieved data at various time intervals. Features of spatial, temporal, and semantic tendency can be used to illustrate these linkages. Semantic and grammatical tendencies in the spoken phrases make up the text modality in addition to the emotional elements present in each modality. Convolutional neural networks

(CNN) and recurrent neural networks (RNN) have been used in deep learning-based models to extract spatial and temporal features in a sequential manner. However, RNNs may not be as good at detecting semantic tendencies in speech as they are at identifying distinct spatial-temporal feature representations. Concurrent spatial-temporal and grammatical (CoSTGA) is a deep learning model that we propose in this paper. It learns spatial, temporal, and semantic representations in parallel in the local feature learning block (LFLB) and fuses them as a latent vector to form an input to the global feature learning block (GFLB).

In this research, we also present the multi-level transformer encoder model (MLTED) and use it to study the performance of multi-level feature fusion against single-level fusion. Utilizing the suggested CoSTGA model, multi-level fusion: first, at the LFLB level, where comparable features (temporal or spatial) are extracted independently from a modality; second, at the GFLB level, where the features of the spatial-temporal combination are fused with the features of the semantic tendency. Dilated causal convolutions (DCC), bidirectional long short-term memory (BiLSTM), transformer encoders (TE), multi-head, and self-attention processes are all included in the proposed CoSTGA model. Utilizing the interactive emotional dyadic motion capture (IEMOCAP) dataset, acoustic and lexical characteristics were retrieved. The weighted and unweighted accuracy, recall, and F1 score achieved by the suggested model are 75.50%, 75.82%, and 75.32%, respectively. These findings suggest that the efficacy and robustness of the model are enhanced by concurrently learning spatial-temporal aspects with semantic inclinations acquired using a multi-level method [1].

A. LIMITATIONS:

- Features from the interactive emotional dyadic motion capture (IEMOCAP) dataset were retrieved, including verbal and auditory aspects. This data collection is inefficient and quite tiny.
- As a result of all these features and processes, the module is difficult for the device to comprehend and process.
- This model uses a lot of processing power and space, but it still achieves weighted and unweighted accuracy of 75.50% and 75.82%, respectively.
- The module's enormous space and time requirements render it immoral to operate.

IV. ARCHITECTURE AND PROPOSED SYSTEM

Our research goal was to create a system that could identify a person's emotions from their speech, which might be useful for a number of AI-related applications. This technology might be employed in smart automobiles to adjust their behavior based on the driver's emotional state, such as slowing down when the driver is furious or afraid, or it could be used in contact centers to play calming music when a caller is upset. Including emotion detection in these kinds of applications might improve user safety and enjoyment. We used the Python Librosa module for audio feature extraction to do this. Librosa is an all-in-one toolkit for audio and music analysis that includes everything needed to create music information retrieval systems. Our main objective was to obtain Mel Frequency Cepstral Coefficients (MFCCs), which are often utilized in speaker recognition and automated speech.

These qualities offer important insights into the voice's spectrum properties. The effect of gender on emotion detection was an intriguing finding in our experiment. We discovered that a 15% improvement in emotion recognition accuracy was achieved by segregating the voices of men and women. This impact may have been caused by gender-specific variations in voice pitch, which affected the outcomes. Numerous feature values were produced by each audio file and were arranged in arrays. Our emotion identification model was then trained using these characteristics in conjunction with the labels given to each emotional category. Managing lost elements in shortened audio files was another difficulty for us. In order to overcome this, we doubled the sampling rate in order to obtain distinct characteristics from every emotional speech sample. A greater sampling frequency might add noise and jeopardize the accuracy of our emotion identification system, so it was crucial to find a compromise. The intricacy and promise of voice emotion identification are demonstrated by this effort, which also emphasizes the need of feature extraction and data pretreatment for reliable outcomes.

A. Data Collection:

The following list of five data sources contains the combined data utilized in this project:
RAVDESS: 2452 audio files, 12 male and 12 female voices; each speaker speaks just two equal-length sentences in each of eight distinct moods, maintaining the lexical aspects (vocabulary) of the utterances constant.

B. PYTHON:

Python is a general-purpose, high-level programming language. Its design concept uses the off-side rule to apply considerable indentation, which promotes code readability.

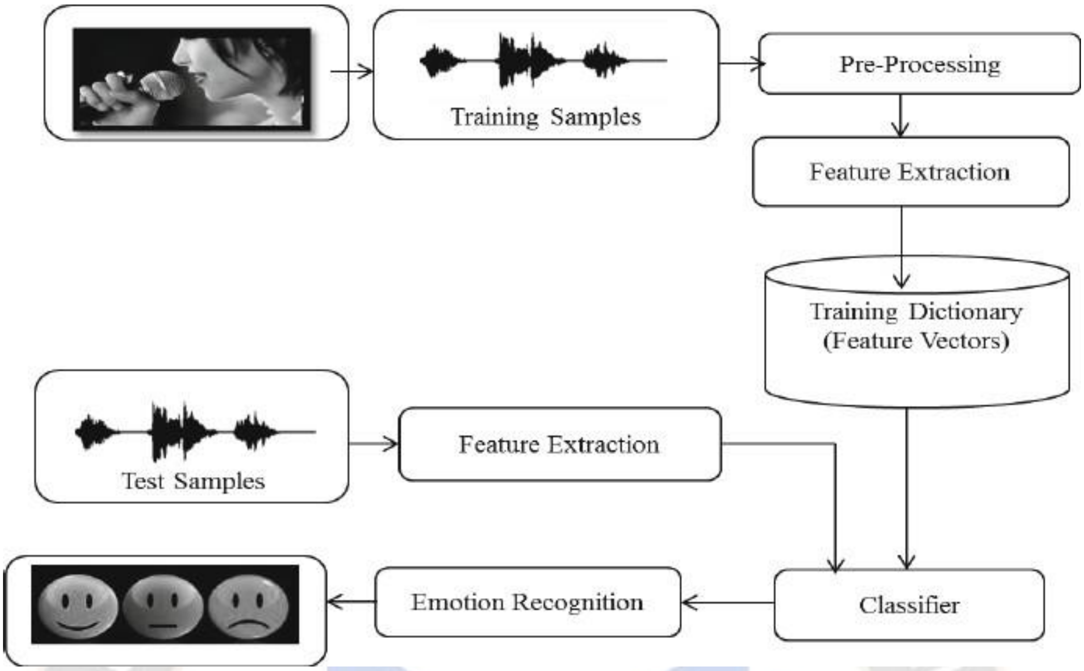


Fig 1: Architectural representation of proposed system

Python uses garbage collection and dynamic typing. It is compatible with several programming paradigms, such as object-oriented, functional, and structured (especially procedural). Because of its extensive standard library, it is frequently referred to as a "batteries included" language. Python was developed by Guido van Rossum in the late 1980s as a replacement for the ABC programming language. Python 0.9.0 was initially made available in 1991. 2000 saw the introduction of Python 2.0. 2008 saw the introduction of Python 3.0, a significant update that was not entirely backwards compatible with previous iterations. The final Python 2 release was Python 2.7.18, which was made available in 2020. Python is one of the most widely used programming languages, period.

C. RAVDESS:

There are 7,356 files in the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS); the total size is 24.8 GB. Twenty-four professional actors—twelve women and twelve men—vocalise two lexically matched phrases in a neutral North American accent for the database. Expressions such as peaceful, joyful, sad, furious, afraid, disgusted, and surprised may all be found in speech, and the same can be found in songs. Every expression has two emotional intensity levels (strong and normal), in addition to a neutral expression. Three modalities are offered for all conditions: Video-only (no sound), Audio-Video (720p H.264, AAC 48kHz,.mp4), and Audio-only (16bit, 48kHz,.wav). Take note that Actor_18 does not have any song files.

D. LibROSA:

Librosa is a useful Python library for sound and music analysis that aids programmers in creating Python programs for handling sound and music document designs. Basically, we use this Python package for sound and music analysis when

we work with sound data, such as in the music era (using Lstm's), Automatic Speech Recognition. The library can handle both traditional and modern tasks related to sound and music processing, and it is easy to use. Under the terms of the ISC License, it is freely available and open source. Time-space sound handling, successive demonstrating, coordinating consonant percussive partition, beat-simultaneous, stacking and translating the sound, symphonious percussive source detachment, conventional spectrogram decay, burden sound from a circle, register of different spectrogram portrayals, and some more are among the elements related to sound records handling and extraction that are supported by the library.

E. KAGGLE:

For those interested in machine learning and data science, Kaggle is an online community platform. Users may use GPU integrated notebooks, search and publish datasets, communicate with other users, and compete with other data scientists to solve data science challenges on Kaggle. With its strong tools and resources, this online platform—which was developed in 2010 by Anthony Goldbloom and Jeremy Howard and purchased by Google in 2017—aims to assist professionals and students in achieving their objectives in the field of data science. On Kaggle as of right now (2021), there are more than 8 million registered users.

F. SYSTEM ARCHITECTURE DIAGRAM:

System architecture diagrams offer a visual representation of the many parts of a system and demonstrate how they interact and communicate with one another. These diagrams show the architecture and structure of a system.

G. MODULES OVERVIEW:

A module is an assembly of build parameters and source files that allows you to segment your project into distinct functional components. There can be one or more modules in your project, and dependencies between modules can exist. Every module can be built, tested, and debugged on its own.

Extra modules can be handy for building code libraries inside your own project or for producing distinct code and resource sets for various device kinds, such phones and wearable’s, while maintaining the same project's scope and sharing some code.

A project's logical division of functionality is called a module. They are mostly employed to improve code maintenance and reusability. Here, three modules are in use.

1) LibROSA:

Librosa is a useful Python library for sound and music analysis that aids programmers in creating Python programs for handling sound and music document designs. Basically, we use this Python package for sound and music analysis when we work with sound data, such as in the music era (using Lstm's), Automatic Speech Recognition.

The library can handle both basic and advanced tasks related to sound and music processing, and it is easy to use. Under the terms of the ISC License, it is freely available and open source.

A few components related to handling and extracting sound recordings are held by the library: beat-simultaneous sound extraction, stacking and translating the sound, Time-space sound handling, successive demonstrating, coordinating consonant percussive partition, burden sound from a circle, register of various spectrogram portrayals, symphonies percussive source detachment, conventional spectrogram decay, and so on.

Librosa uses several sign handling techniques to separate the components from the sound signs and helps visualize them.

It will help us to carry out:

- Sound sign analysis related to music. The library offers its master clients, who may be interested in managing sound records, a great deal of flexibility.
- Make reference to the use of standard procedures. It provides the structural building pieces needed to create frameworks for the recovery of music data.
- The fundamental components of music data recovery (MIR).

2) RAVDESS:

First off, the RAVDESS has 7356 clips, compared to many sets' less than 200 clips. S1 and S2 Figs show the factorial design of the RAVDESS. Each of the 24 professional actors that make up the RAVDESS performs 104 distinct vocalizations representing a range of emotions, including joyful, sad, angry, afraid, surprised, disgusted, calm, and neutral. Given that imaging studies have demonstrated that important brain areas become used to repeated exposures of the same stimulus, this variety may prove helpful in repeated measurements designs.

Machine learning researchers can also benefit from having a huge corpus of recordings. The verified database offers a big set for training and testing various algorithms, making it especially well-suited to machine learning techniques including supervised learning, such emotion classifiers.

There are two emotional intensity degrees for each emotion: normal and intense. Only two other sets that we are aware of offer a controlled modification of intensity.

One of the most noticeable characteristics of emotion is intensity, which is central to a number of theories of emotion. Please take note that in these works, the words "intensity" and "activation" have been used interchangeably. Intensity frequently functions as one of multiple orthogonal axes in a multidimensional emotional space in these models. Vibrant facial and verbal emotions are perceived and recognized more reliably than their less intense equivalents.

Additionally, observers are more likely to identify extreme facial emotions than less intense ones, and they also respond to intense facial imitation. Thus, when researchers are looking for distinct, unmistakable emotional exemplars, powerful displays could be helpful.

On the other hand, when examining minute variations in emotional perception or when researchers are looking for realistic representations, normal intensity expressions could be necessary.

3) KAGGLE:

For those interested in machine learning and data science, Kaggle is an online community platform. Users may use GPU integrated notebooks, search and publish datasets, communicate with other users, and compete with other data scientists to solve data science challenges on Kaggle. With its strong tools and resources, this online platform—which was developed in 2010 by Anthony Goldbloom and Jeremy Howard and purchased by Google in 2017—aims to assist professionals and students in achieving their objectives in the field of data science. On Kaggle as of right now (2021), there are more than 8 million registered users.

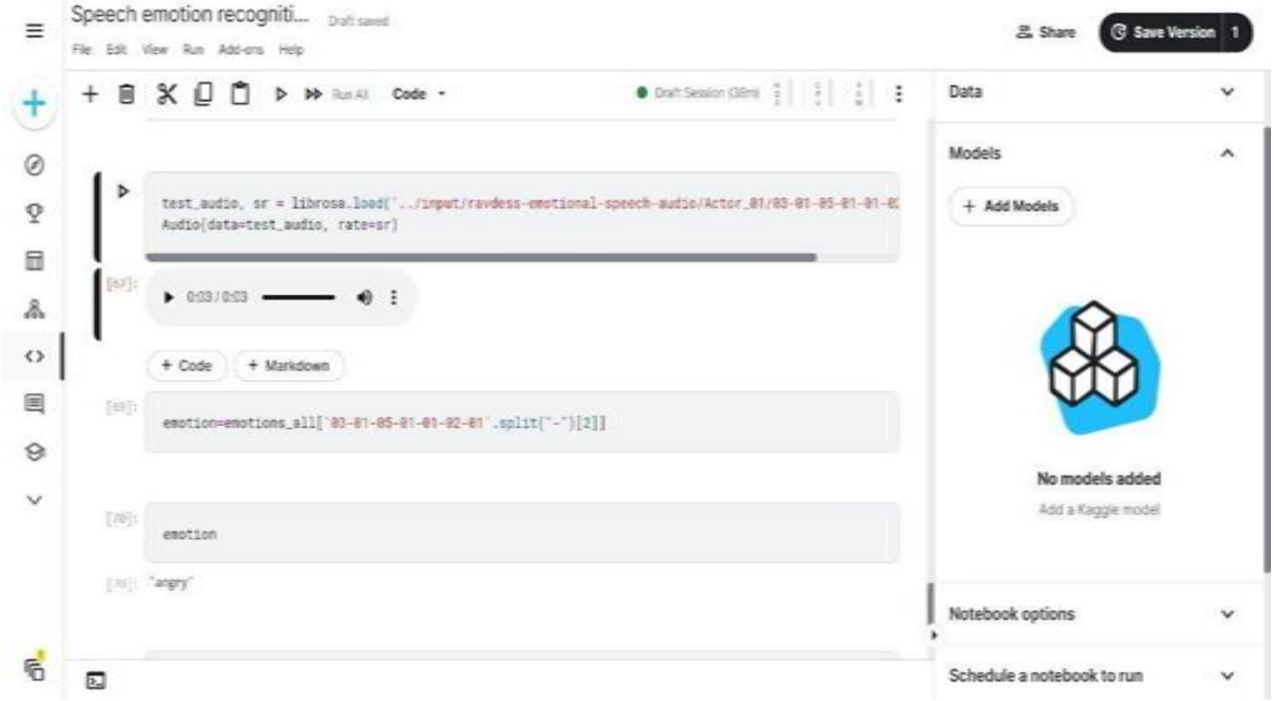
Competitions are one of the sub-platforms on Kaggle that have contributed to its popularity. For data scientists, "Kaggle Competitions" hold great significance, much like HackerRank does for software developers and computer engineers. You can read more about them in our guide to Kaggle Competitions and get step-by-step instructions on how to analyze a dataset in our Kaggle Competition Tutorial. Companies and organizations offer a large number of difficult data science jobs with high payouts in data science contests like Kaggle's or DataCamp's, where data scientists of all expertise levels compete to complete the tasks. Additionally, Kaggle offers the Kaggle Notebook, which functions similarly to DataCamp Workspace in that it lets you update and execute your code for data science challenges directly from your browser, saving you the trouble of setting up a new development environment or relying only on your local computer to handle the heavy work.

V. RESULTS AND DISCUSSION:

Our study's findings support the efficacy of the speech analysis-based emotion identification system we suggested. We were able to extract Mel Frequency Cepstral Coefficients (MFCCs), an important feature in voice and speaker recognition, by using the Librosa package for feature extraction. Our emotion categorization model was built on the vital spectrum information these MFCCs supplied. One interesting discovery was that separating the sample into male and female voices significantly increased the accuracy of emotion identification. This gender-based division resulted in a 15% improvement in our model's accuracy. This finding implies that gender variations in voice pitch and other acoustic characteristics may have an impact on how emotions are categorized.

The problem of managing shorter audio samples with absent characteristics was another task we took on. These shorter utterances allowed us to extract more unique characteristics by increasing the sample rate by two. This method kept the identification system's quality while increasing the feature set's comprehensiveness without appreciably increasing noise.

In the area of emotion identification, our model performed competitively. Our dataset yielded an accuracy rate of 80%, indicating the ability of the suggested system to recognize human emotions from auditory signals.



```
test_audio, sr = librosa.load('../input/ravdess-emotional-speech-audio/Actor_01/03-01-05-01-01-0
Audio(data=test_audio, rate=sr)
```

```
[57]: ▶ 0:03 / 0:03
```

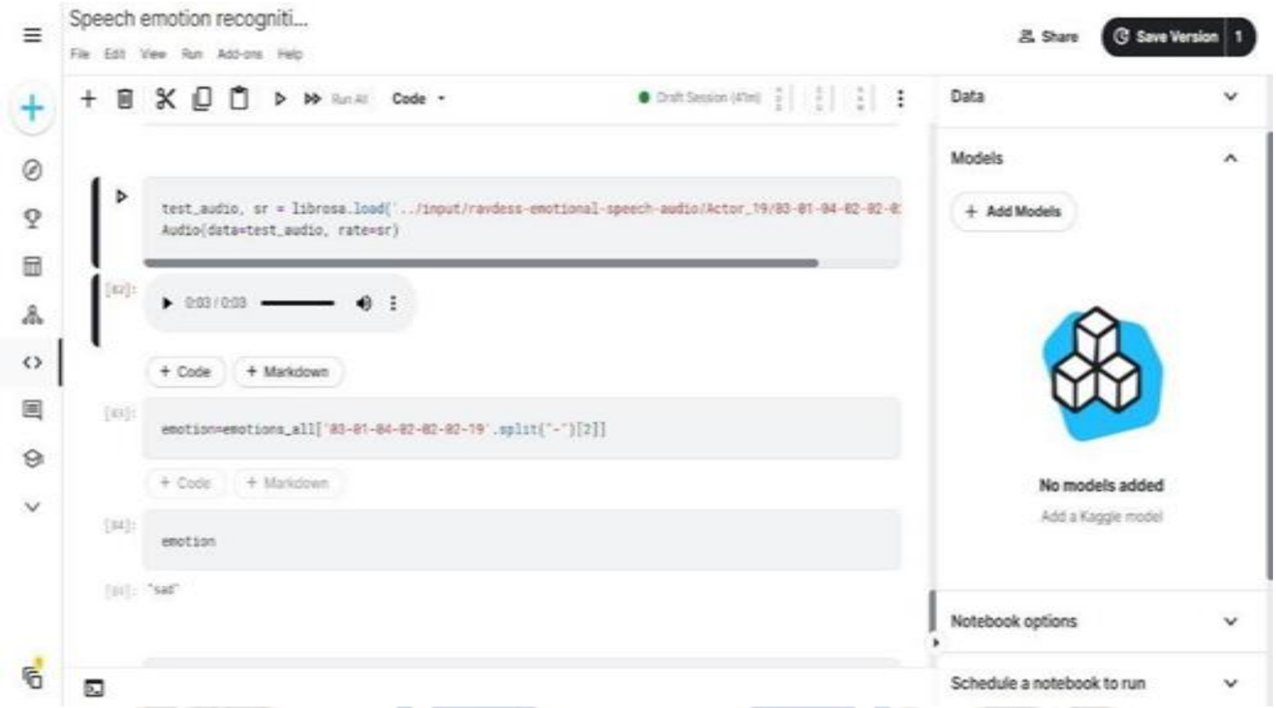
```
[69]: emotion=emotions_all['03-01-05-01-01-02-01'.split("-")[2]]
```

```
[70]: emotion
```

```
[70]: 'angry'
```

Fig2. Represents the Output of the proposed model (sample1).

DOI: https://doi.org/10.17762/ijritcc.v11i10.8472
***Article Received:** 26 July 2023 **Revised:** 12 September 2023 **Accepted:** 01 October 2023*



```
test_audio, sr = librosa.load('../input/ravdess-emotional-speech-audio/Actor_19/03-01-04-02-02-0
Audio(data=test_audio, rate=sr)
```

```
[82]: 0:03 / 0:03
```

```
[83]: emotion=emotions_all['03-01-04-02-02-02-19'.split("-")[2]]
```

```
[84]: emotion
```

```
[84]: 'sad'
```

Fig3.Represents the Output of the proposed model (sample1).

Our study has important ramifications for speech emotion identification systems and offers various new insights. Above all, our results validate the idea that gender-specific models might improve the identification accuracy of emotions. This discovery is consistent with the well-established variations in voice features and pitch between males and females. More specialized and accurate applications may result from the capacity to take gender into account while recognizing emotions, particularly in situations where gender affects how emotions are expressed. A key component of our study is the higher sampling rate, which enhanced feature capture in shorter audio samples. By using this technique, we may improve the resilience of the emotion identification system and reduce data loss. While raising the sample rate, it's important to maintain equilibrium because oversampling might result in the introduction of noise. Thus, future studies might investigate the ideal sample rate for various auditory circumstances and emotional settings. Our approach's practicality in real-world applications is demonstrated by our promising overall accuracy of 80% on the dataset. The practical value of our study is shown by the prospective use cases of this technology, which include driving safety in smart vehicles and improving customer service in contact centers. This emphasizes how crucial it is to carry out further research and development in the area of speech emotion identification since it has potential uses in the business and in safety-critical settings. Our research highlights the potential applications of voice-based emotion identification across several disciplines. Emotion extraction and analysis from speech has the potential to improve user experience, have a big influence on human-computer interaction, and even be safe and healthy. Through the resolution of issues related to feature extraction, gender-based modeling, and data preparation, we have made great progress in developing emotion detection systems. To reach the full potential of speech emotion identification in real-world applications, more study can build on and enhance these findings.

VI. CONCLUSION:

We have identified the optimal CNN model for our emotion classification task after developing a large number of alternative models and evaluating each one independently. Overall, the suggested model's accuracy is 77.57%. Additional data to work with might improve the performance of our model. When the RAVDESS dataset was utilized, the model much improved. One of the main contributors to the accuracy and simplicity factors in the dataset is the 2452 audio files, which have 12 male and 12 female speakers.

The model did a great job of differentiating between male and female voices, which presents us with a new challenge for the module. With the dataset's assistance, the distinguishing voices may be categorized more quickly and easily. In comparison to its predecessors, the proposed Speech Emotion Recognition System is much simpler and has superior accuracy overall. To ensure the viability of the SER System, the time and space complexity were also reduced to a minimum.

ACKNOWLEDGMENT

No funding sources.

AUTHOR CONTRIBUTION

Author 1 implemented the concept specified by the author 2 under the supervision of authors 3 & 4. The authors 3 & 4 & 5 drafted the article under the guidance of author 2.

REFERENCES

[1] Samuel Kakuba, Alwin Poulose & Dong Seog Han Deep Learning - Based Speech Emotion Recognition UsingnMulti - Level Fusion of Concurrent Features

[2] Bagus Tris Atmaja, Akira Sasou. Speech Emotion and Naturalness Recognitions with Multitask and Single-Task Learnings (IEEE-2022)

[3] Chenghao Zhang. Autoencoder With Emotion Embedding for Speech Emotion Recognition. (IEEE-2021)

[4] Jennifer Santoso, Takeshi Yamada, Kenkichi Ishizuka. Speech Emotion Recognition Based on Self-Attention Weight Correction for Acoustic and Text Features. (IEEE-2022)

[5] Ting-Wei Sun. EndtoEnd Speech Emotion Recognition with Gender Information (IEEE-2020)

[6] Xiaohan Xia, Dongmei Jiang, Hichem Sahli. Learning Salient Segments for Speech Emotion Recognition Using Attentive Temporal Pooling (IEEE-2020)

[7] Felicia Andayani, Lau Bee Theng, Mark Teekit Tsun, Caslon Chua. Hybrid LSTM-Transformer Model for Emotion Recognition from Speech Audio Files. (IEEE-2022)

[8] Karam Kumar Sahoo, Ishan Dutta, Muhammad Fazal Ijaz, Marcin Woźniak, Pawan Kumar Singh. TLEFuzzyNet: Fuzzy Rank-Based Ensemble of Transfer Learning Modelsfor Emotion Recognition from Human Speeches. (IEEE-2021)

[9] Shunzhi Yang, Zheng Gong, Kai Ye. EdgeRNN A Compact Speech Recognition Network with Spatio-Temporal Features for Edge Computing. (IEEE-2021)

[10] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri. A Comprehensive Review of Speech Emotion Recognition Systems. (IEEE- 2021)

[11] Jorge Oliveira, Isabel Praça. On the Usage of Pre- Trained Speech Recognition Deep Layers to Detect Emotions. (IEEE- 2021)

[12] Danai Styliani Moschona. An Affective Service based on Multi-Modal Emotion Recognition, using EEG enabled Emotion Tracking and Speech Emotion Recognition. (IEEE-2020)

[13] Ryota Sato;Ryohei Sasaki;Norisato Suga;Toshihiro Furukawa. Creation and Analysis of Emotional Speech Database for Multiple Emotions Recognition. (IEEE-2020)

[14] Misaki Sakurai;Tetsuo Kosaka. Emotion Recognition Combining Acoustic and Linguistic Features Based on Speech Recognition Results. (IEEE-2021)

[15] Yuanchao Li;Peter Bell;Catherine Lai. Fusing ASR Outputs in Joint Training for Speech Emotion Recognition. (IEEE-2022).