ANALYSING SPEECH EMOTION USING NEURAL NETWORK ALGORITHM

Valavala Surya Teja[1], Kodi Sravya[2], Kattoju Jahnavi Annapurneswari[3], Shaik Sajid[4], Undrajavarapu Likhita[5], Arepalli Rajesh[6]

[1]CST, Sri Vasavi Engineering College(A), Pedatadepalli,Tadepalligudem-534101
[2]CST, Sri Vasavi Engineering College(A), Pedatadepalli,Tadepalligudem-534101
[3]CST, Sri Vasavi Engineering College(A), Pedatadepalli,Tadepalligudem-534101
[4]CST, Sri Vasavi Engineering College(A), Pedatadepalli,Tadepalligudem-534101
[5]CST, Sri Vasavi Engineering College(A), Pedatadepalli,Tadepalligudem-534101
[6]Sr. Assistant Professor, Department of CSE, Sri Vasavi Engineering College(A),Pedatadepalli, Tadepalligudem-534101

---***---

Abstract - *In this paper we propose a deep learning model to identify speech emotion. This paper presents a comprehensive study on the application of neural network algorithms for improving SER performance. We propose a novel approach that leverages deep learning techniques, including convolutional neural networks and LSTM, to extract and analyse acoustic features from speech signals. The results demonstrate the effectiveness of our model in accurately identifying emotional states from speech, showcasing its potential for real-world applications. Our findings contribute to the growing body of knowledge in the field of human-computer interaction, sentimental analysis, healthcare and highlight the potential of neural networks in enhancing the accuracy and robustness of speech emotion recognition systems.*

***Key Words*: Speech Emotion Recognition, Deep Learning, Datasets: RAVDESS, SAVEE, CREMA-D, TESS, Acoustic Features, Neural Network Model**

1.INTRODUCTION

Understanding and interpreting emotions in human speech is a fascinating and challenging aspect of human-computer interaction and artificial intelligence. The ability to recognize emotions from spoken words has significant implications in various fields, from improving customer service to aiding individuals with emotional or communication difficulties. Advancements in deep learning techniques, particularly Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks, have offered promising avenues for enhancing the accuracy and efficiency of speech emotion analysis. These neural network models have demonstrated remarkable capabilities in learning intricate patterns and representations from audio data, making them well-suited for the task of speech emotion recognition.

This paper delves into the realm of speech emotion analysis using CNN and LSTM models. We explore the challenges and complexities associated with understanding emotions in the spoken word and investigate how neural networks can be employed to tackle these challenges. Our research focuses on the development, training, and evaluation of a neural network model dedicated to speech emotion analysis. In this paper, we will present our approach, discuss our methodologies, and provide insights into how CNN and LSTM models can be harnessed to improve the accuracy and robustness of speech emotion recognition. By shedding light on the potential of these models in analysing speech emotions, our work aims to contribute to the broader field of affective computing, offering practical solutions for emotion-aware applications and services.

1.1 Benefits:

Human-Computer Interaction: Emotion recognition in speech can be used to make human-computer interaction more natural and intuitive. It can be applied in virtual assistants, chatbots, and other AI-driven systems to respond more empathetically to user emotions.

Market Research and Customer Feedback: Businesses can use speech emotion analysis to gauge customer sentiment and emotional reactions in real-time, providing valuable insights for improving products and services.

Healthcare: In healthcare, analysing speech can help in early detection of mood disorders, such as depression and anxiety. It can also be used to monitor and provide care for patients with cognitive or emotional issues.

Customer Service: In customer support, speech emotion analysis can be used to monitor the emotions of both customers and support agents, ensuring better service quality and training opportunities.

Accessibility: Speech emotion analysis can enhance accessibility features for differently-abled individuals. For example, it can assist those with visual impairments to perceive emotional cues during interactions.

2. LITERATURE SURVEY

The literature review stands out as a crucial phase in the software development process. It entails an exploration of prior research conducted by various authors in the relevant field. We will analyze and build upon key articles to enhance our work.

2.1 Li-Min Zhang suggested a research paper on speech emotion recognition, they proposed a novel algorithm called F-Emotion to select speech emotion features and established a parallel deep learning model to recognize different types of emotions. He concluded that F-Emotion algorithm significantly improves speech emotion recognition accuracy.

2.2 Mohammad Reza In this paper, a speech emotion recognition system based on a 3D CNN is suggested to analyse and classify the emotions. He concluded, the three-dimensional reconstructed phase spaces of the speech signals were calculated in order to recognise the emotion in speech.

2.3 Lei Yang suggested a research paper on Speech Emotion Analysis of Netizens using Machine learning approach in 2021.The algorithm used in the study is referred to as the "PSOGA-CDBN (PGCDBN) model. He concluded that this hybrid deep learning model is designed for speech emotion recognition (SER) and achieved an average recognition accuracy .

2.4 Reem Hamed Aljuhani suggested a research paper on Speech Emotion Analysis using Machine learning approach in 2021.The algorithm used in the study is support vector machine (SVM). He concluded that only four emotions were studied: anger, happiness, sadness, and neutral.

2.5 Yogesh Kumar suggested a research paper on Speech Emotion Analysis using Machine learning approach in 2019.He used Convolutional Neural Network (CNN) algorithm. He concluded that only three emotions were detected: anger, sadness, and neutral.

3. EXISTING SYSTEM

There are several existing systems and approaches for analysing speech emotion, each with its own strengths and weaknesses. Some of the commonly used methods and systems include:

3.1 Limited Emotion Recognition Accuracy: Even with advances in machine learning and deep learning, emotion recognition accuracy is not always optimal. Understanding and categorizing complex human emotions from speech remains a challenging task, especially when dealing with subtle or mixed emotions.

- **3.2 Lack of Standardized Datasets:** The availability of high-quality, diverse, and large emotion-labelled speech datasets is limited. The quality and quantity of training data can significantly impact the performance of these systems.

- **3.3 Contextual Understanding:** Current systems often lack a deep understanding of the context in which speech occurs. Context can heavily influence the interpretation of emotional content, but most systems do not incorporate this contextual information.

- **3.4 Handling Ambiguity:** Speech can be ambiguous, and the same acoustic features may represent different emotions in various contexts. Existing systems may struggle with disambiguating emotional cues.

- **3.5 Real-time Processing:** Some applications, like real-time voice assistants, require immediate emotion recognition. Existing systems may not be optimized for real-time processing, leading to delays in responses.

- **3.6 Bias and Fairness:** Emotion analysis systems can be susceptible to biases present in the training data, leading to potential fairness issues and misclassification of emotions for certain demographic groups

4. PROPOSED WORK

Proposed system in the context of analysing speech emotion typically refers to a conceptual framework or technology designed to recognize and categorize emotions conveyed through spoken language. Such a system can have various components and features. Here's an outline of what a proposed system for analysing speech emotion might include:

- **High-Quality and Diverse Datasets:** A critical foundation for any speech emotion analysis system is a comprehensive, high-quality dataset that covers a wide range of emotions and demographic groups. Collecting diverse data can help reduce biases and improve the system's generalizability.

- **Emotion Intensity and Continuous Monitoring:** Consider the ability to measure the intensity of emotions and track emotional changes over time, especially in applications like mental health monitoring.

- **Interpretability:** Develop models that provide explanations for their emotion recognition decisions. Interpretability is crucial for user trust and for understanding why a particular emotion was recognized.

- **Validation and Evaluation:** Continuously evaluating the performance of a speech emotion analysis system is a fundamental and ongoing process that underpins its reliability and relevance. The aim is to not only ensure the accuracy and effectiveness of the system but also to fine-tune its capabilities and address potential shortcomings

- **Scalability and Adaptability:** Designing a system that can be easily scaled to handle larger datasets and adapted to new domains or languages is a strategic approach to ensure the long-term versatility and robustness of a speech emotion analysis system. This design flexibility not only accommodates the growing data needs of the system but also facilitates its expansion into new contexts.

METHODOLOGY:

From figure 4.2 we can observe the step by step process of the proposed model.

Step 1: Data Collection:

We have gathered data from datasets named RAVDESS, SAVEE, CREMA-D, and TESS, where emotions are explicitly labelled and associated with each piece of recorded speech.

From figure 4.1 we can observe that the graph shows the count of emotions we took.

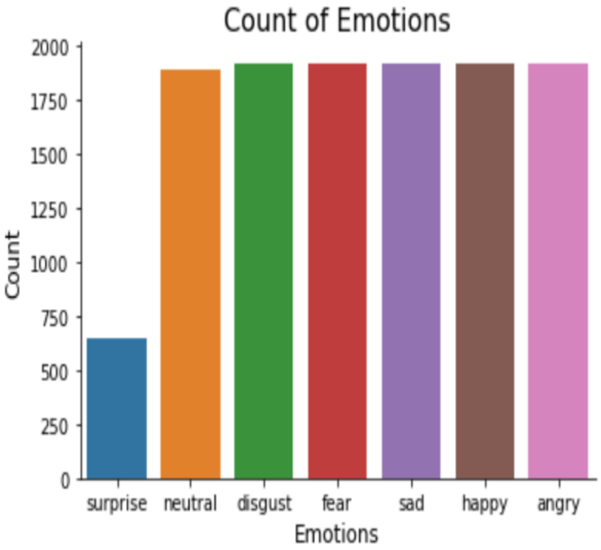


Figure 4.1

Step 2: Data Preprocessing:

Extracting relevant features from the audio, such as MFCCs (Mel-frequency cepstral coefficients). These preprocessing steps play a pivotal role in ensuring that the collected data is ready for effective model training and subsequent deployment. Splitting the data into training, validation, and testing sets.

Step 3: Model Training and Validation:

Continuously training and fine-tuning many models, aiming to improve their performance based on evaluation standards, is the path to finding the most suitable model for your specific task or challenge. This process involves trying out various techniques and configurations. To achieve the highest accuracy and effective generalization on the validation data, this iterative approach entails making changes to hyperparameters like learning rates, regularization strengths, and model architectures.

Step 4: Model Evaluation:

Accuracy serves as a valuable measure for assessing a model's overall performance due to its simplicity and widespread use. This metric is particularly common in binary classification tasks, making it easy to grasp and communicate. Nevertheless, it's vital to acknowledge its constraints, as accuracy can be misleading when dealing with imbalanced datasets.

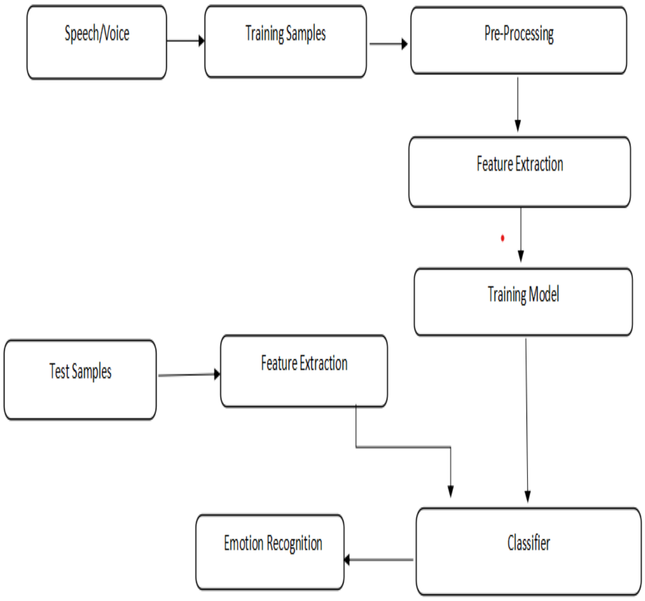


Figure 4.2

5. Dataset

Dataset Url:

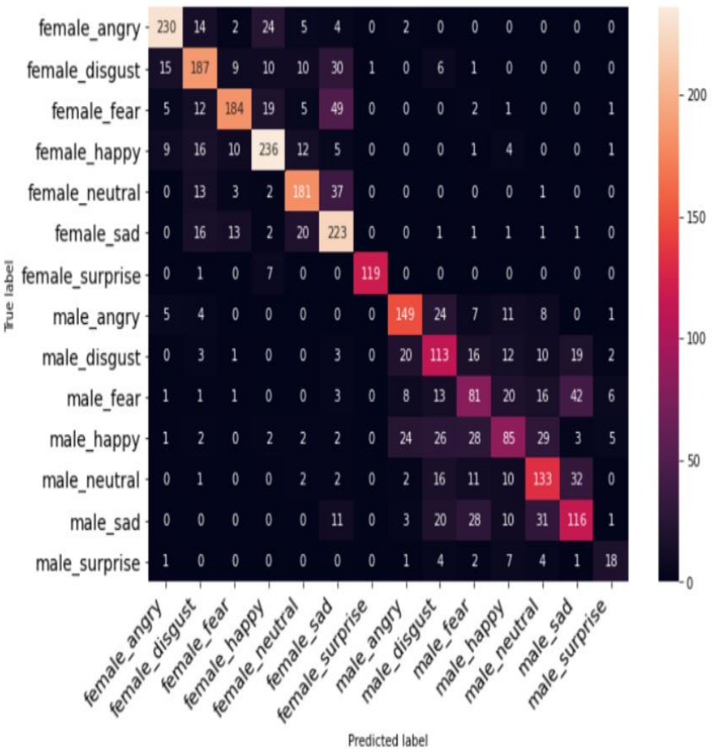
https://www.kaggle.com/datasets/uwrfkaggler/ravdess-emotional-speech-audio

https://www.kaggle.com/datasets/barelydedicated/savee-database

https://www.kaggle.com/datasets/ejlok1/toronto-emotional-speech-set-tess

https://www.kaggle.com/datasets/ejlok1/cremad

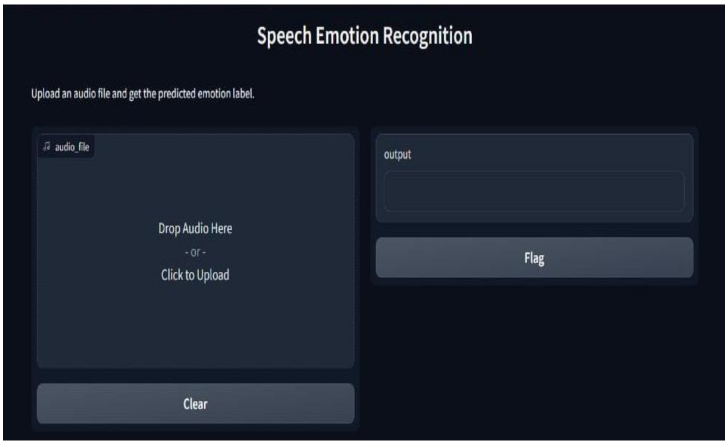
These datasets, RAVDESS, SAVEE, CREMA-D, and TESS, are integral to the fields of speech and emotion recognition. RAVDESS offers extensive audio and video recordings of actors portraying various emotions, while SAVEE features British English speakers expressing basic emotions. CREMA-D is distinctive for its audio and video recordings of actors displaying emotions, and TESS focuses on North American English speakers' emotional speech. Researchers and machine learning practitioners frequently rely on these datasets to train models for emotion recognition and speech processing, driving progress in understanding and utilizing emotional cues in human communication.



	female_angry	female_disgust	female_fear	female_happy	female_neutral	female_sad	female_surprise	male_angry	male_disgust	male_fear	male_happy	male_neutral	male_sad	male_surprise
female_angry	230	14	2	24	5	4	0	2	0	0	0	0	0	0
female_disgust	15	187	9	10	10	30	1	0	6	1	0	0	0	0
female_fear	5	12	184	19	5	49	0	0	0	2	1	0	0	1
female_happy	9	16	10	236	12	5	0	0	0	1	4	0	0	1
female_neutral	0	13	3	2	181	37	0	0	0	0	0	1	0	0
female_sad	0	16	13	2	20	223	0	0	1	1	1	1	1	0
female_surprise	0	1	0	7	0	0	119	0	0	0	0	0	0	0
male_angry	5	4	0	0	0	0	0	149	24	7	11	8	0	1
male_disgust	0	3	1	0	0	3	0	20	113	16	12	10	19	2
male_fear	1	1	1	0	0	3	0	8	13	81	20	16	42	6
male_happy	1	2	0	2	2	2	0	24	26	28	85	29	3	5
male_neutral	0	1	0	0	2	2	0	2	16	11	10	133	32	0
male_sad	0	0	0	0	0	11	0	3	20	28	10	31	116	1
male_surprise	1	0	0	0	0	0	0	1	4	2	7	4	1	18

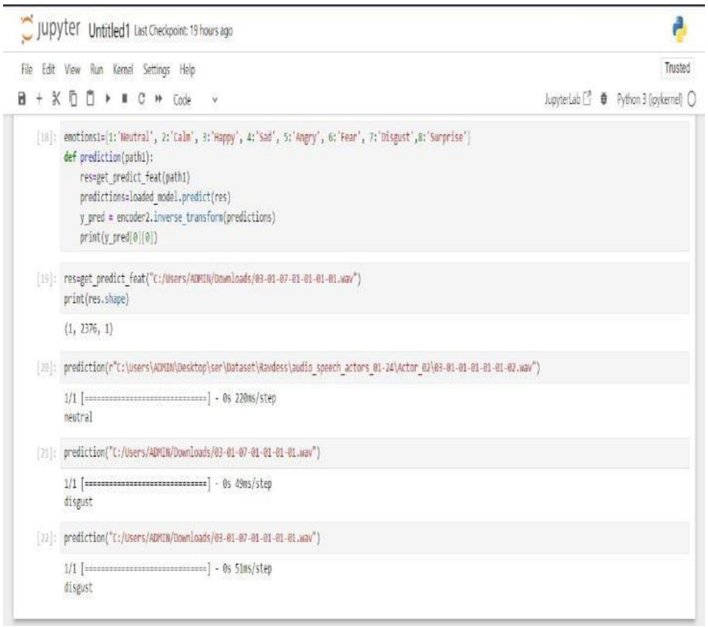
Figure 6.1: Evaluated results of the model using CNN-LSTM

6. User Interface



Upload an audio file and get the predicted emotion label.

Figure 6.2



jupyter Untitled1 Last Checkpoint: 19 hours ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)

```
[18]: emotions1={1:'Neutral', 2:'Calm', 3:'Happy', 4:'Sad', 5:'Angry', 6:'Fear', 7:'Disgust',8:'Surprise'}
      def prediction(path1):
          res=get_predict_feat(path1)
          predictions=loaded_model.predict(res)
          y_pred = encoder2.inverse_transform(predictions)
          print(y_pred[0][0])

[19]: res=get_predict_feat("C:/Users/ADMIN/Downloads/03-01-07-01-01-01-01.wav")
      print(res.shape)

      (1, 2376, 1)

[20]: prediction(r"C:\Users\ADMIN\Desktop\ser\Dataset\Ravdess\audio_speech_actors_01-24\Actor_02\03-01-01-01-01-01-02.wav")

      1/1 [==============================] - 0s 220ms/step
      neutral

[21]: prediction("C:/Users/ADMIN/Downloads/03-01-07-01-01-01-01.wav")

      1/1 [==============================] - 0s 49ms/step
      disgust

[22]: prediction("C:/Users/ADMIN/Downloads/03-01-07-01-01-01-01.wav")

      1/1 [==============================] - 0s 51ms/step
      disgust
```

Figure 6.3

7. CONCLUSION

In this paper we used special computer techniques to understand how people feel when they talk. This paper made the computer better at figuring out emotions in speech, and it can be used in many things like talking with computers, checking how someone feels, and more. We used clever tools like NLP, CNN, and LSTM, and looked at big sets of speech data like RAVDESS, SAVEE, CREMA-D, and TESS. This helps us understand emotions better when people talk. This paper shows how technology can be more like us, making our conversations with computers more friendly and helpful.

7. ACKNOWLEDGEMENT

We are really grateful to **Arepalli Rajesh, M.Tech, Sr.Assistant Professor**, Department of Computer Science and Engineering, for all of his help and assistance during the creation of this project. His meticulous critiques insightful observations, and continuous encouragement have been crucial in helping to shape and bring our research to completion. In addition to helping our project succeed, his commitment to creating an atmosphere of academic rigor and creative thinking has given us a love for lifelong learning and discovery. We deeply appreciate him for his perseverance, awareness, and dedication, which have served as the cornerstones of our path.

We also extend our appreciation to Dr.D.Jaya Kumari and Dr. G.V.N.S.R. RatnaKara Rao for cooperation and encouragement.

Lastly, we acknowledge the unwavering support of our staff of the Department of Computer Science and Engineering played a crucial role in facilitating our project work, and we thank them for their contributions.

REFERENCES

[1] Li-Min Zhang, Giap Weng, Yu-Beng Leau and Haoyan- “A Parallel-Model Speech Emotion Recognition Network Based on Feature Clustering” in 2023.

[2] Mohammad Reza Falah Zadeh, Edris Zaman Farsa, Ali Harimi, Arash Ahmadi and Ajith Abraham-“3D Convolutional Neural Network for Speech Emotion Recognition With Its Realization on Intel CPU and NVIDIA GPU” in 2022.

[3] Lei Yang, Kai Xie, Chang Wen and Jian-Biao- “Speech Emotion Analysis of Netizens Based on Bidirectional LSTM and PGCDB” in 2021.

[4] Reem Hamed Aljuhani, Areej Alshutayri and Shahd Alahdal- “Arabic Speech Emotion Recognition from Saudi Dialect Corpus” in 2021.

[5] Dr. Yogesh Kumar, Dr. Manish Mahajan- “Machine Learning Based Speech Emotions Recognition System” in 2019.

[6] Neethu Sundarprasad- “Speech Emotion Detection Using Machine Learning Techniques” in 2018.

[7] Kumari S and Perinban D - “Speech Emotion Recognition Using Machine Learning” in 2021.

[8] T. Sai Samhith and G. Nishika- “Speech Emotion Recognition Using Machine Learning Algorithms” in 2021.

[9] S.G. Shaila and A. Sindhu- “Speech Emotion Recognition Using Machine Learning Approach” in 2022.

[10] Husbaan I. Attar and Nilesh K. Kadole- “Speech Emotion Recognition System Using Machine Learning” in 2022.

[11] Nithya Roopa and Prabhakaran- “Speech Emotion Using Deep Learning” in 2018.

[12] Shubham Singh Chaudhary and Sachine Garg- “Speech Emotion Recognition” in 2021.

[13] Hima Keerthi Sagiraju and Pritam Sharma- “Speech Emotion Recognition Using Machine Learning” in 2021.

[14] Madhusudhan and Mahesh Kumar- “Speech Emotion Recognition” in 2021.

BIOGRAPHIES

Description “**Valavala Surya Teja** pursuing third year in the department of Computer Science and Technology, at Sri Vasavi Engineering College, Tadepalligudem, affiliated with JNTU Kakinada, AP.The joy of discovering new insights and the satisfaction of applying them in real-life situations fuel my passion for lifelong learning."

Description “**Kodi Sravya** pursuing third year in the department of Computer Science and Technology, at Sri Vasavi Engineering College, Tadepalligudem, affiliated with JNTU Kakinada, AP. I prioritize active listening as a key skill, paired with a quick learning aptitude, propelling me to excel in dynamic and fast-paced environments."

Description “**Kattoju Jahnavi Annapurneswari** pursuing third year in the department of Computer Science and Technology, at Sri Vasavi Engineering College, Tadepalligudem, affiliated with JNTU Kakinada, AP. I am confident in my ability to cultivate a successful career, approaching my work with both zeal and meticulous attention."

Description “**Shaik Sajid** pursuing third year in the department of Computer Science and Technology, at Sri Vasavi Engineering College, Tadepalligudem, affiliated with JNTU Kakinada, AP. I prioritize active listening as a key skill, paired with a quick learning aptitude, propelling me to excel in dynamic and fast-paced environments."

Description "**Undrajavarapu Likhita** pursuing third year in the department of Computer Science and Technology, at Sri Vasavi Engineering College, Tadepalligudem, affiliated with JNTU Kakinada, AP. With a resolute mindset, I tackle tasks by immersing myself in the subject matter, swiftly grasping complexities, and successfully completing assigned work with determination."