Real-Time Speech Emotion Recognition Using Machine Learning

Ms. P. Bhavani Durga[1], Mr. N. V Ratna Kishor Gade[2], Dr. D. Jaya Kumari[3]

[1]M. Tech, Student, Department of Computer Science, Sri Vasavi Engineering College(A), Pedatadepalli, Tadepalligudem – 534101.
[2]Assistant Professor, Department of CSE, Sri Vasavi Engineering College(A), Pedatadepalli, Tadepalligudem – 534101.
[3]Professor & HOD, Department of CSE, Sri Vasavi Engineering College(A), Pedatadepalli, Tadepalligudem – 534101.

Abstract:
This project is centred around examining spoken language emotion, a critical aspect of human-computer interaction and artificial intelligence. Leveraging advanced techniques such as Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) models, we aim to decipher and understand the nuanced expressions of emotions conveyed through speech. The investigation draws from diverse datasets, including RAVDESS, SAVEE, TESS, and CREAM-D, each carefully chosen to encapsulate a broad spectrum of emotions and real-world speech scenarios. The project unfolds by elucidating the methodologies employed in dataset selection, the prepossessing of raw audio data, and the intricacies of utilizing CNN and LSTM techniques for speech emotion analysis. The main objective is to create a robust and reliable outcome model capable of accurately classifying and interpreting emotions in speech across various contexts. Practical applications of this Research extend to fields such as sentiment analysis and potential contributions to mental health monitoring. Through this project, we contribute valuable insights to the evolving landscape of speech emotion analysis, addressing the inherent challenges and exploring opportunities for enhancing human-computer interaction and emotional intelligence in artificial systems.

Keywords: Speech, CNN and LSTM, Voice, RAVDESS, SAVEE, TESS, and CREAM-D.

1. INTRODUCTION

Speech Emotion Recognition (SER) refers to identifying emotions conveyed in spoken language. It entails utilizing DL methods to examine audio signals and extract features that reveal the underlying emotional content of speech. The primary goal of SER is to automatically detect emotions such as happiness, sadness, anger, fear, and more from speech recordings.

This field has numerous applications, including—but not limited to—enhancing user experience in chatbots, virtual assistants, or human-computer interfaces by making systems more responsive to user emotions. CNNs, RNNs, and their variants, like Long-Short-Term Memory (LSTM) networks, are commonly employed for their ability to learn intricate patterns in speech data.

We analyze emotions in speech using advanced techniques like convolutional neural networks (CNN) and long short-term memory (LSTM) models. We've chosen four diverse datasets—RAVDESS, SAVEE, TESS, and CREAM-D—to cover various emotions and speech scenarios. The RAVDESS dataset combines genuine and acted emotions, forming a solid foundation for our study. Additionally, SAVEE, TESS, and CREAM-D datasets contribute real-world and actor-induced emotions, ensuring our model can handle various emotional contexts in speech.

In the upcoming sections, we'll discuss how we selected and prepared our datasets and the details of using CNN and LSTM techniques. Our goal is to better understand emotional cues in speech and apply this knowledge practically, such as in human-computer interaction and potentially in mental health monitoring. This project aims to contribute insights to the growing field of speech emotion analysis.

2. LITERATURE SURVEY

The most crucial step in the software development process is the literature review. This will describe some preliminary research that was carried out by several authors on this appropriate work. We are going to consider some critical articles and further extend our work.

Lei Yang, Kai Xie, Chang Wen, and Jian-Biao[1] Proposed a research article, "Speech Emotion Analysis of Netizens Based on Bidirectional LSTM and PGCDBN," in IEEE 2021. They suggested a research paper on Speech Emotion Analysis of Netizens using a Machine learning approach in 2021. The algorithm used in the study is called 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.

Li-Min Zhang, Giap Weng, Yu-Beng Leau and Haoyan[2] Proposed a research article"A Parallel-Model Speech Emotion Recognition Network Based on Feature Clustering" IEEE 2023. They suggested a research paper on speech emotion recognition, they suggested a new algorithm named F-Emotion for choosing features related to speech emotions and created a parallel deep learning model to identify various emotions. He concluded that the F-Emotion algorithm significantly improves speech emotion recognition accuracy.

Mohammad Reza Falah Zadeh, Edris Zaman Farsa, Ali Harimi, Arash Ahmadi and Fajita Abraham [3] Proposed a research article, "3D Convolutional Neural Network for Speech Emotion Recognition With Its Realization on Intel CPU and NVIDIA GPU" IEEE 2022. This paper suggests a speech emotion recognition system based on a 3D CNN to analyze and classify emotions. He concluded that the three-dimensional reconstructed phase spaces of the speech signals were calculated to recognize the emotion in speech.

Jianhua Zhang, Zhong Yin, and Peng Chen[4]Proposed a research article, "Emotion Recognition Methods Based on multi-channel EEG Signals", as well as multi-modal physiological signals are reviewed. According to the standard pipeline for emotion recognition, we review different feature extraction.

Jeevan Singh Deusi and Elena Irena Popa [5] Proposed a research article. Studies propose enhancements to feature selection from speech signals and pattern recognition algorithms for recognizing emotions. The system utilized to conduct the experiments is Emotion Pi, using the Raspberry Pi 3 B+.

Sajad A. Shah and Sahilpreet Singh[6]Proposed a research article. Research indicates room for improvement in selecting features from speech signals and employing pattern recognition algorithms for emotion recognition. The system utilized to conduct the experiments is currently being utilized. Emotion Pi, using the Raspberry Pi 3 B+

3. EXISTING METHODOLOGY

The current system for analyzing speech emotion has a few common issues. It often needs help to detect emotions, making it less reliable and accurate. Additionally, it faces challenges when dealing with different accents and languages, leading to potential misinterpretations. Another concern is the system's difficulty in capturing subtle emotional nuances, making it less effective in understanding the finer aspects of human expression. These limitations highlight the need for improvements in emotion analysis technology to enhance its overall performance and reliability.

Limitations of the Existing System:

1. Struggles with diverse accents and dialects, causing misclassifications.
2. Often optimized for one language, it is less effective in multilingual settings.
3. Lacks data for subtle or less common emotions.
4. Limited real-time capabilities for applications like virtual assistants.

4. PROPOSED SYSTEM

The proposed system aims to analyze speech emotion using NLP and DL techniques. It involves processing speech data to extract relevant features like pitch, tone, and intensity. These features are then used as input to ML algorithms, such as neural networks or support vector machines, to classify the emotional content of the speech into different categories, like happiness, sadness, anger, etc.

Advantages of the Proposed System:

1. Enhanced Accuracy: Neural networks (e.g., CNNs, RNNs) extract features from raw speech, yielding superior emotion classification.
2. Automated Feature Learning: Neural networks learn features, reducing manual engineering and ensuring adaptability to diverse linguistic and emotional variations.
3. Effective Generalization: Neural networks generalize patterns for real-world SER, recognizing emotions across contexts and user groups. It involves processing speech data to extract relevant features like pitch, tone, and intensity. These features are then used as input to ML algorithms, such as neural networks or SVMs, to classify the emotional content of the speech into different categories, like happiness, sadness, anger, etc.

5. DATA SET DESCRIPTION

The dataset collected from Kaggle [13]

The RAVDESS dataset, the Ryerson Audio-Visual Database of Emotional Speech and Song, is a comprehensive resource for SER projects. It includes audio clips of actors and actresses simulating a wide range of emotions, making it a valuable tool for training models to recognize emotions accurately. The RAVDESS dataset features diverse emotions such as happiness, sadness, anger, and fear and includes different intensity levels, enabling researchers to study nuanced emotional expressions. The CREMA-D dataset, the Crowd-Sourced Emotional Multi-modal Actors Dataset, is another significant resource for SER. CREMA-D contains emotional speech recordings from various actors and is unique in its multi-modal approach, incorporating audio and video data. This dataset is valuable for understanding the relationship between facial expressions and vocal cues when expressing emotions, enriching the study of human emotional expression. The SAVEE dataset, the Surrey Audio-Visual Expressed Emotion database, is a smaller but still essential dataset for SER. It features emotional speech samples with a focus on the core emotions: happiness, sadness, anger, and fear. While smaller in scale than other datasets, SAVEE remains a valuable resource for developing and testing emotion recognition models. The TESS dataset, the Toronto Emotional Speech Set, is notable for its focus on acted emotional speech, emphasizing the emotional expressions of actors. It features seven different emotions and is particularly useful for understanding and recognizing emotional expressions in speech, as it offers various acted emotional states.

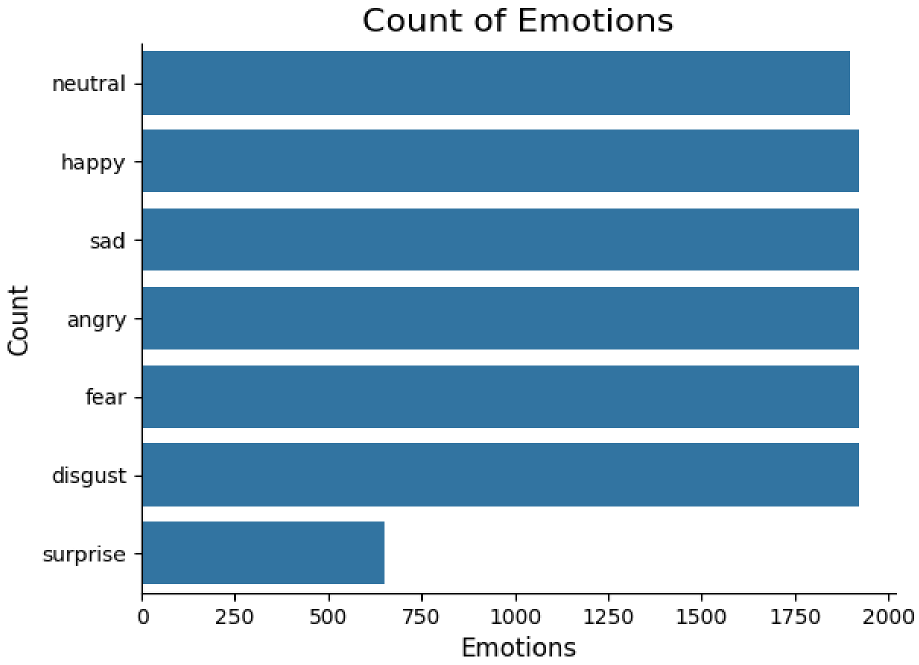


Figure 1:DATA SET

6. DESIGN & METHODOLOGY

Convolutional Neural Networks(CNN): It is employed to learn hierarchical representations of acoustic features from speech signals, enhancing the accuracy of emotion classification.

Long Short-Term Memory (LSTM): These networks capture temporal dependencies in acoustic features, allowing for effective modelling of sequential information and improved emotion classification.

Data Collection:

Gather a dataset of speech samples labelled with corresponding emotions (e.g., happiness, sadness, anger, etc.).

Preprocessing:

Convert speech signals to numerical representations (e.g., MFCC features).

Normalize the features to ensure consistency.

Feature Extraction:

Extract relevant features from the preprocessed speech signals (e.g., Mel-Frequency Cepstral Coefficients (MFCCs), prosodic features, etc.).

Model Training:

I am dividing the dataset into training, validation and testing.

Train machine learning models (e.g., Support Vector Machines (SVM), Random Forests, Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), etc.) on the training data.

Model Evaluation:

Evaluate the trained models using the validation set.

Fine-tune hyperparameters to improve performance.

Real-time Processing:

Implement a real-time processing pipeline to capture and process incoming speech signals. Then, apply the trained model to classify emotions in real time.

Feedback Loop:

Collect feedback from the real-time system to improve performance.

Adapt the model based on user interaction and real-world usage.

Deployment:

Deploy the real-time speech emotion recognition system in the desired application or platform (e.g., virtual assistants, customer service systems, etc.).

Monitoring and Maintenance:

Monitor system performance over time.

Perform periodic maintenance and updates to ensure continued effectiveness.

7. SYSTEM DESIGN

A system architecture is a conceptual model that defines a system's structure, behaviour, and views.

Audio Input:

This is the source of the speech signals. It could be a microphone capturing real-time speech or prerecorded audio files.

Preprocessing:

Audio signals often need preprocessing before feature extraction. This could entail activities like minimizing noise, resampling, and normalization.

Feature Extraction:

Extract relevant features from the preprocessed audio signals. Commonly used features for speech emotion recognition include Mel-frequency Cepstral coefficients(MFCCs), pitch, energy, and spectral features.

Machine Learning Model:

Train a machine learning model using labelled datasets. Popular choices include Support Vector Machines (SVM), Random Forests, or deep learning models such as Convolutional Neural Networks (CNNs) or Recurrent Neural Networks (RNNs).

Real-Time Inference:

Deploy the trained model for real-time inference. The model predicts the emotional state based on the extracted features as new audio data comes in.

Emotion Output:

Display or log the predicted emotions. This output could be visualized in real time or stored for further analysis.

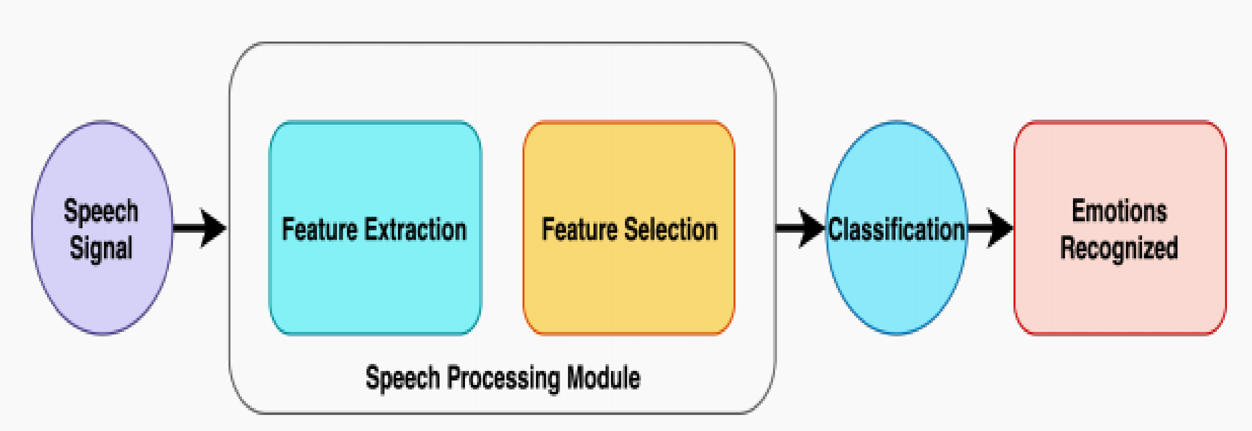


Figure 2: SYSTEM DESIGN

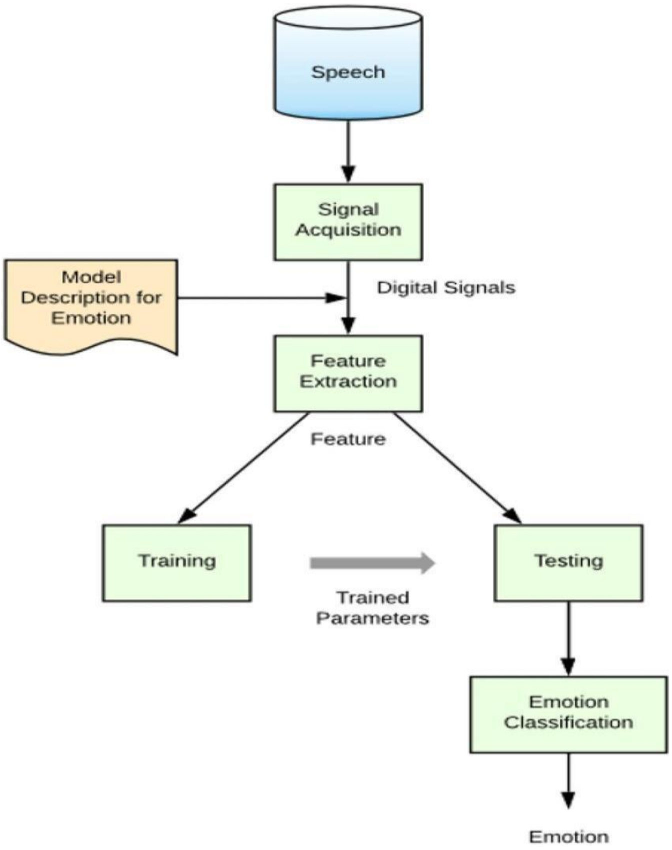
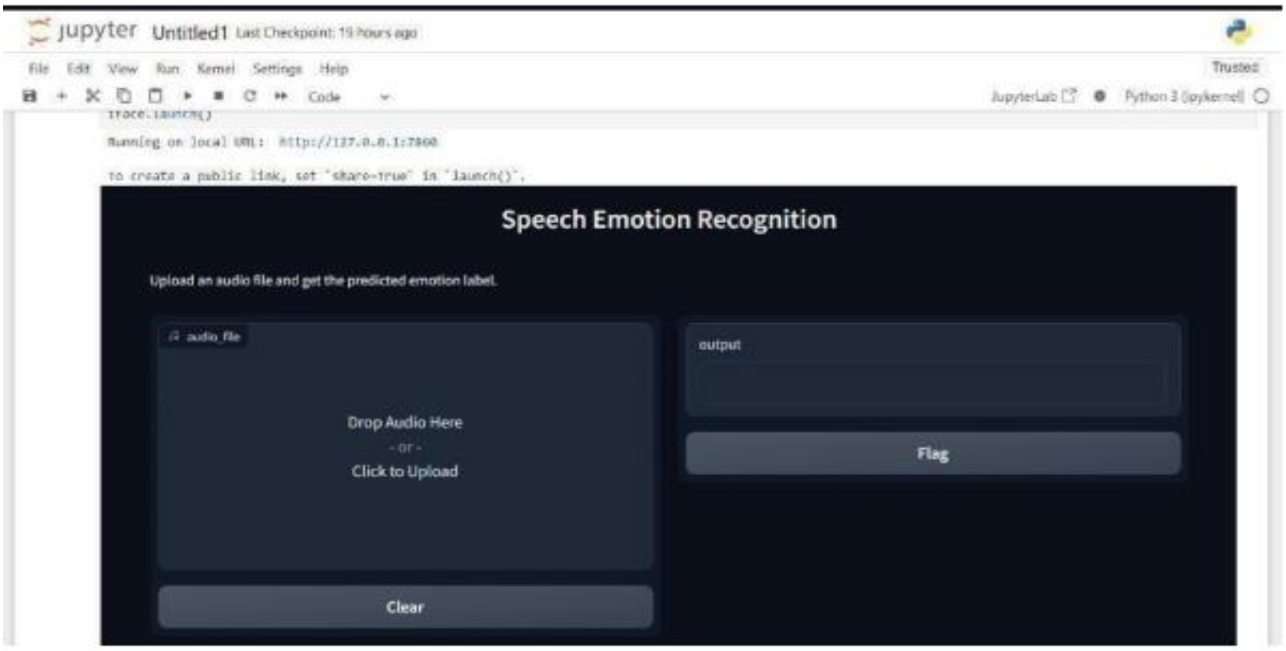


Figure 3: System Architecture

8. RESULTS

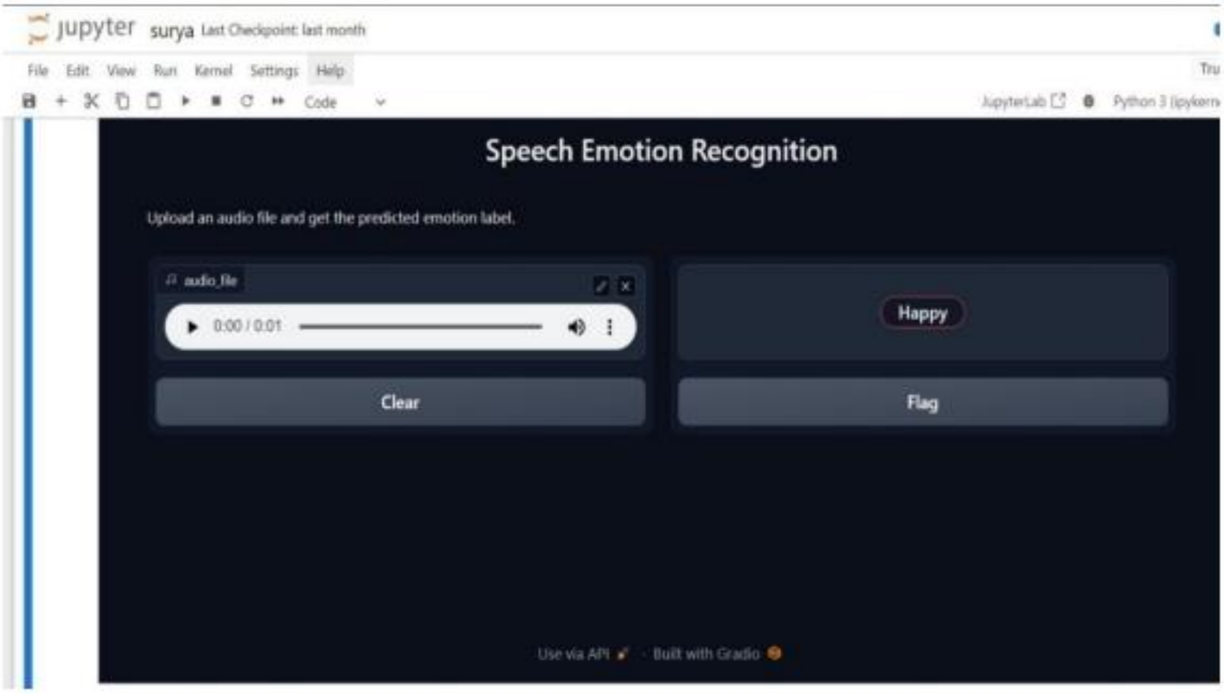

jupyter Untitled1 Last Checkpoint: 19 hours ago
File Edit View Run Kernel Settings Help
Trusted
Code
JupyterLab
Python 3 (ipykernel)

```
Iface.launch()
Running on local URL:  http://127.0.0.1:7860

To create a public link, set `share=True` in `launch()`.
```

Speech Emotion Recognition

Upload an audio file and get the predicted emotion label.


jupyter surya Last Checkpoint: last month
File Edit View Run Kernel Settings Help
Tru
Code
JupyterLab
Python 3 (ipykern

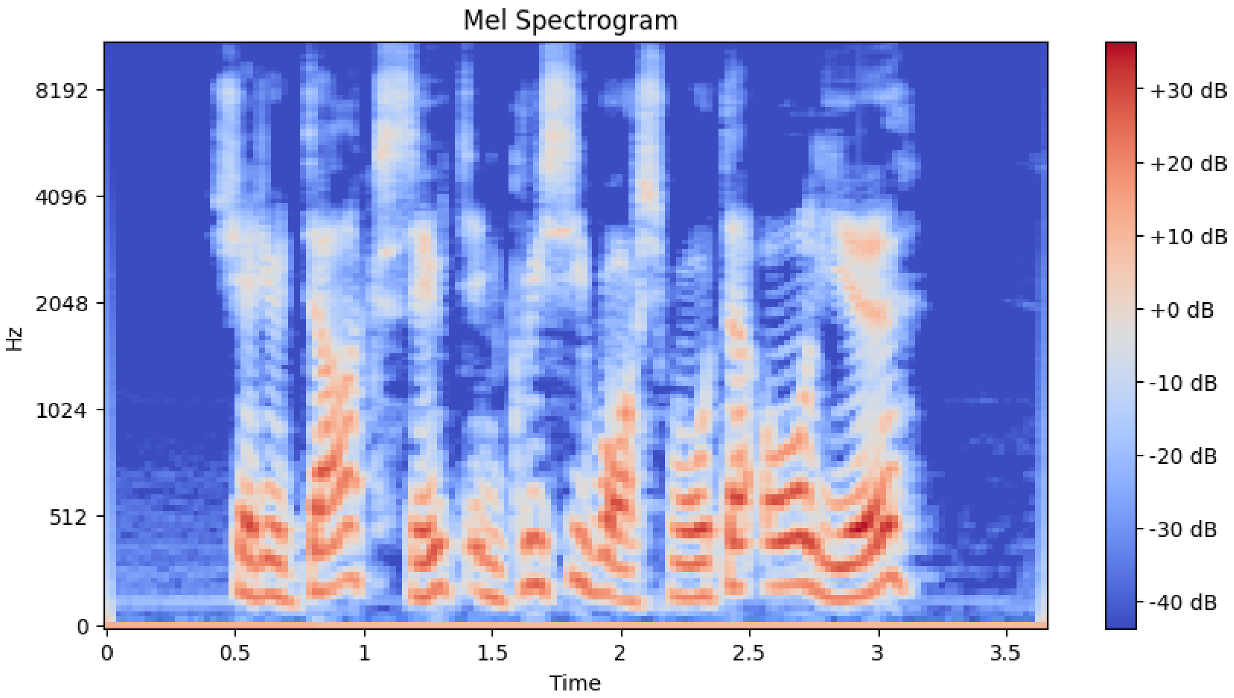
Speech Emotion Recognition

Upload an audio file and get the predicted emotion label.

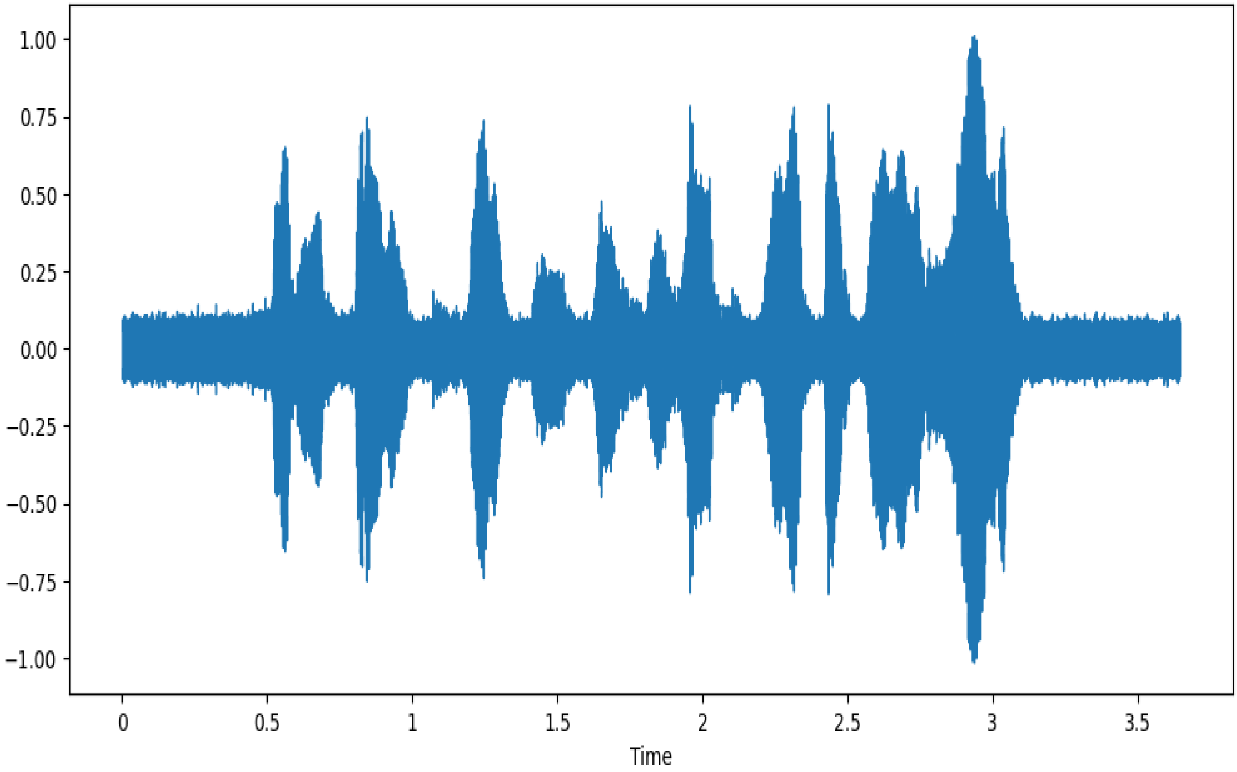
audio_file
0:00 / 0:01
Happy
Clear
Flag
Use via API · Built with Gradio

Speech Emotion Recognition

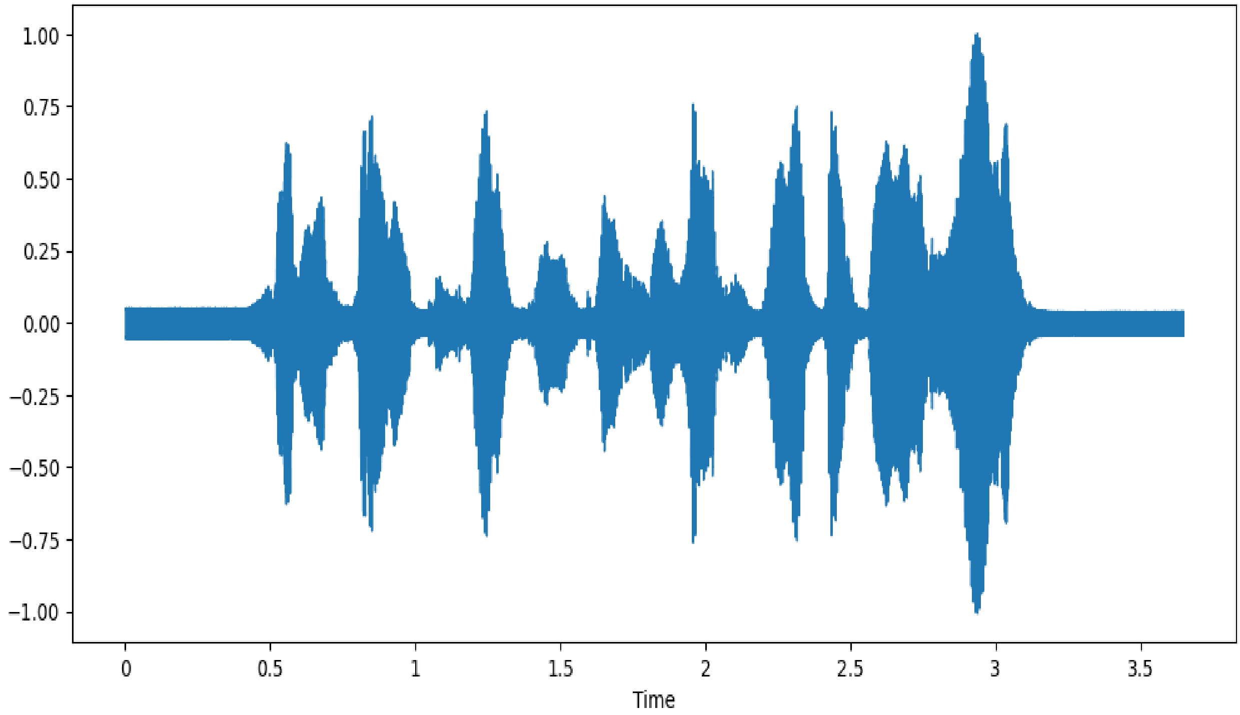
Explanation: From the above window, we can see the input is taken as one sample audio file is uploaded, and that output is happy in the above image.



Mel spectrogram of an audio signal



Audio with time



Audio with noise

9. CONCLUSION

In this project, we used special computer techniques to understand how people feel when they talk. This project 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 speech data sets like RAVDESS, SAVEE, CREMA-D, and TESS. This helps us understand emotions better when people talk. This project shows how technology can be more like us, making our conversations with computers more friendly and helpful.

10. FUTURE SCOPE

In future SER work, areas of exploration include cross-lingual and cross-cultural recognition, real-time applications, emotion intensity estimation, privacy and ethics, adaptive context-aware models, and benchmark dataset development. These endeavours seek to enhance the precision and relevance of emotion recognition, making it adaptable to diverse linguistic and cultural contexts while addressing ethical considerations and fostering human-computer interaction advancements.

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. Arijit, Dey., Soham, Chattopadhyay., Pawan, Kumar, Singh., Ali, Ahmadian., Massimiliano, Ferrara., Ram, Sarkar. (2020). A Hybrid Meta-Heuristic Feature Selection Method Using Golden Ratio and Equilibrium Optimization Algorithms for Speech Emotion Recognition.IEEE Access, doi: 10.1109/ACCESS.2020.3035531
10. Shiqing, Zhang., Shiliang, Zhang., Tiejun, Huang., Wen, Gao. (2018). Speech Emotion Recognition Using Deep Convolutional Neural Network and Discriminant Temporal Pyramid Matching. IEEE Transactions on Multimedia, doi: 10.1109/TMM.2017.2766843
11. M.Shamim, Hossain., Ghulam, Muhammad. (2019). An Audio-Visual Emotion Recognition System Using Deep Learning Fusion for a Cognitive Wireless Framework. IEEE Wireless Communications, doi: 10.1109/MWC.2019.1800419
12. Mehmet, Bilal, Er. (2020). A Novel Approach for Classification of Speech Emotions Based on Deep and Acoustic Features. IEEE Access, doi: 10.1109/ACCESS.2020.3043201
13. 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