SPEECH EMOTION RECOGNITION

Understanding Emotions Through Voice

Presented By: Group G

Aruna Gurung
Adriana Penaranda
Carlos Rey Pinto
Pujan Shrestha
Haldo Jose Somoza

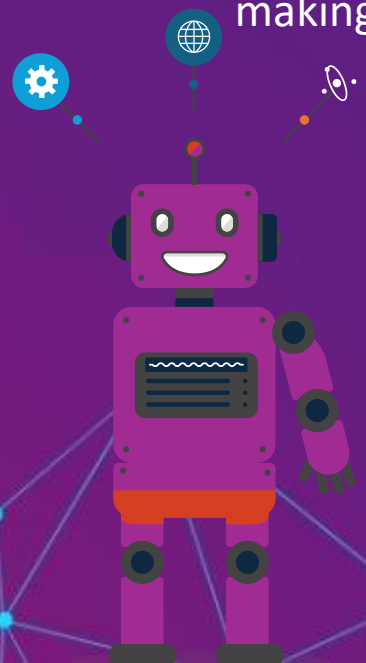
BDM 3035 - Big Data Capstone Project

TABLE OF CONTENTS

INTRODUCTION

PROBLEM STATEMENT

The main challenge in Speech Emotion Recognition (SER) is accurately identifying and understanding human emotions from speech. Emotions are expressed through changes in tone, pitch, and rhythm, which can vary a lot between different people and situations, making it hard to detect emotions consistently.



SIGNIFICANCE AND RELEVANCE

Speech Emotion Recognition (SER) is crucial for improving human-computer interaction, especially in areas like customer service. By detecting emotions in real-time, companies can better respond to customer needs, leading to improved satisfaction.

PROJECT GOAL

The project goal is to create a user-friendly prototype of a machine learning model that accurately detects emotions from speech and integrate it into an application, with the aim of applying this system in customer service settings to improve real-time interactions with users.

DATA COLLECTION AND PREPARATION

DATA OVERVIEW

DATA SOURCE

RAVDESS speech-only dataset with 1,440 audio files from 24 actors (12 female, 12 male), each expressing emotions like calm, happy, sad, angry, fearful, surprise, and disgust at two intensities, plus neutral expressions. The actors vocalize two neutral North American statements, providing a robust foundation for emotion recognition model training.

PORPUSE

Provides a strong foundation for training and evaluating the emotion recognition model, ensuring it generalizes well to new data.

DATA COLLECTION AND PREPARATION

DATA PREPROCESSING STEPS

01

FUTURE EXTRACTION

The initial step in data preprocessing was to extract meaningful features from the raw audio files using the Librosa library.

•**Key Features Extracted:**

1.Mel-Frequency Cepstral Coefficients (MFCCs): Capture the power spectrum of the audio. Essential for identifying the timbral qualities of speech, which are often linked to emotion.

2.Chroma Features: Reflect the tonal content of the audio. Capture the harmonic characteristics, which can vary with different emotions.

3.Spectrograms: Visual representations of the frequency spectrum as it varies over time.

```python
#DataFlair - Extract features (mfcc, chroma, mel) from a sound file
def extract_feature(file_name, mfcc, chroma, mel):
    with soundfile.SoundFile(file_name) as sound_file:
        X = sound_file.read(dtype="float32")
        sample_rate=sound_file.samplerate
        result=np.array([])
        if chroma:
            stft=np.abs(librosa.stft(X))
        if mfcc:
            mfccs=np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=40).T, axis=0)
            result=np.hstack((result, mfccs))
        if chroma:
            chroma=np.mean(librosa.feature.chroma_stft(S=stft, sr=sample_rate).T,axis=0)
            result=np.hstack((result, chroma))
        if mel:
            #mel=np.mean(librosa.feature.melspectrogram(X, sr=sample_rate).T,axis=0)
            mel=np.mean(librosa.feature.melspectrogram(y=X, sr=sample_rate).T,axis=0)
            result=np.hstack((result, mel))
    return result
```

DATA COLLECTION AND PREPARATION

DATA PREPROCESSING STEPS

02 DATA SPLITTING

To ensure proper model evaluation and prevent overfitting:

- **Training Set:** For learning emotion patterns.
- **Validation Set:** For fine-tuning and preventing overfitting.
- **Test Set:** For unbiased final performance evaluation.

03 DATA BALANCING

It was used SMOTE (Synthetic Minority Over-sampling Technique) to balance the dataset by generating synthetic samples for underrepresented emotions

During training, the model performed better without SMOTE, suggesting that synthetic oversampling was not beneficial. It was reverted to the original class distribution and focused on optimizing the model using the naturally occurring data.

Challenges Encountered

- Data Balancing: SMOTE introduced noise; reverted to original data distribution.
- Overfitting: Managed through cross-validation to ensure model generalization.
- Audio Processing: Complex task due to diverse formats and limited tools.

METHODOLOGY

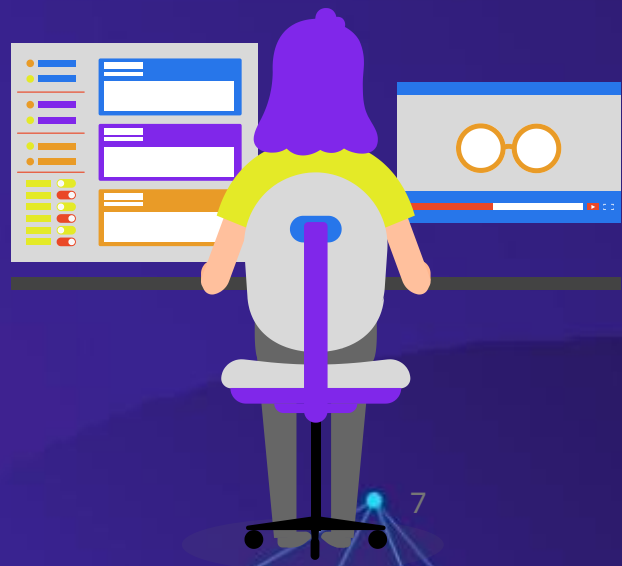
Machine Learning Algorithms and Analytical Technique

Primary Model: Multi-Layer Perceptron (MLP)

- A type of artificial neural network consisting of multiple layers of neurons (nodes).
- Includes an input layer, one or more hidden layers, and an output layer.
- Hidden layers capture non-linear relationships within the data, essential for distinguishing between different emotional states in speech patterns.

Advanced Techniques Used:

- GridSearchCV: Used for hyperparameter tuning, systematically searches for the optimal configuration to maximize model performance.
- Cross-Validation: Employed to assess the model's robustness using k-fold cross-validation, ensuring the model generalizes well on unseen data.



METHODOLOGY

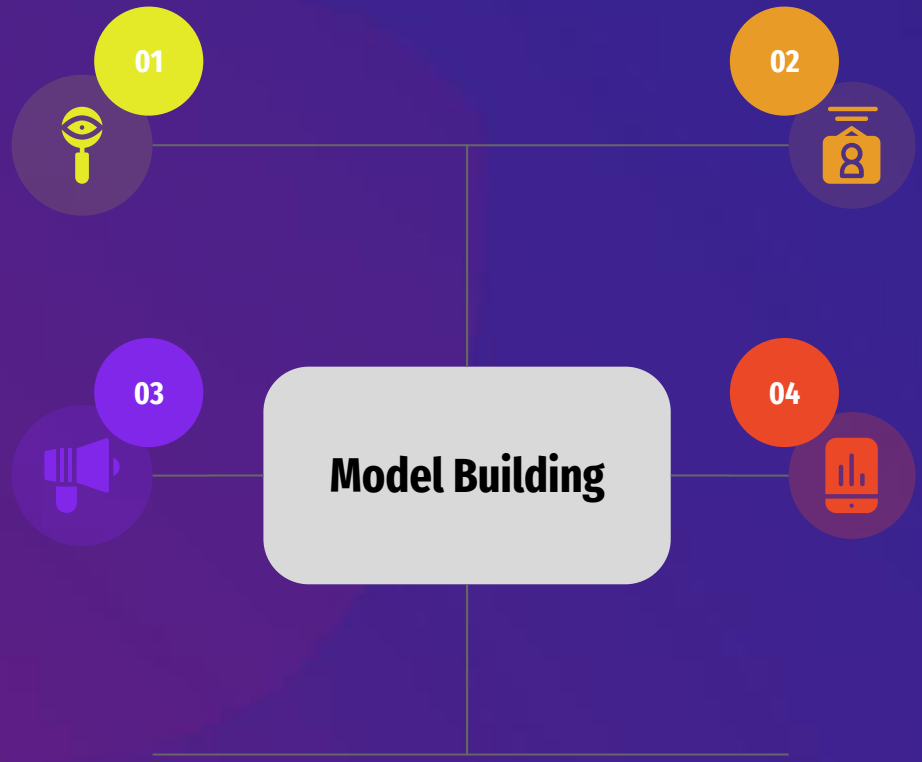
MLP Training, Validation, and Optimization Process

Trainning Process

The model was trained on MFCCs and chroma features, with iterative weight adjustments to reduce prediction error.

Model Evaluation

Validated on a separate dataset not seen during training.



Performance Metrics:

Accuracy: Proportion of correctly predicted emotions.
Precision, Recall, F1-Score: Additional evaluation to understand performance across underrepresented emotional categories.

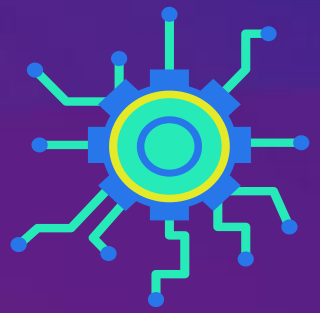
Optimization Techniques:

GridSearchCV: Tuning key hyperparameters such as regularization term (alpha), number of hidden layers, and neurons

Final Goal: Balance model complexity with generalization ability, avoiding overfitting.

METHODOLOGY

Justification for Using MLP and Associated Techniques

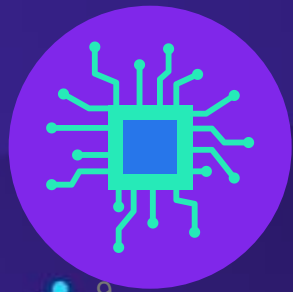


Why MLP?

- **Captures Complex Patterns:** MLP is adept at handling non-linear and subtle patterns in speech, which are essential for accurately recognizing emotions that may not be straightforward.

- **Flexible Architecture:** MLP allows customization of layers and neurons, enabling the model to be tailored to the specific characteristics of the dataset, thereby enhancing its effectiveness in emotion recognition.

- **Comparison with Linear Models:** Unlike linear models, which may struggle with capturing higher-order interactions between features, MLP excels in identifying complex relationships between input variables and output labels, making it more suitable for tasks like emotion recognition.

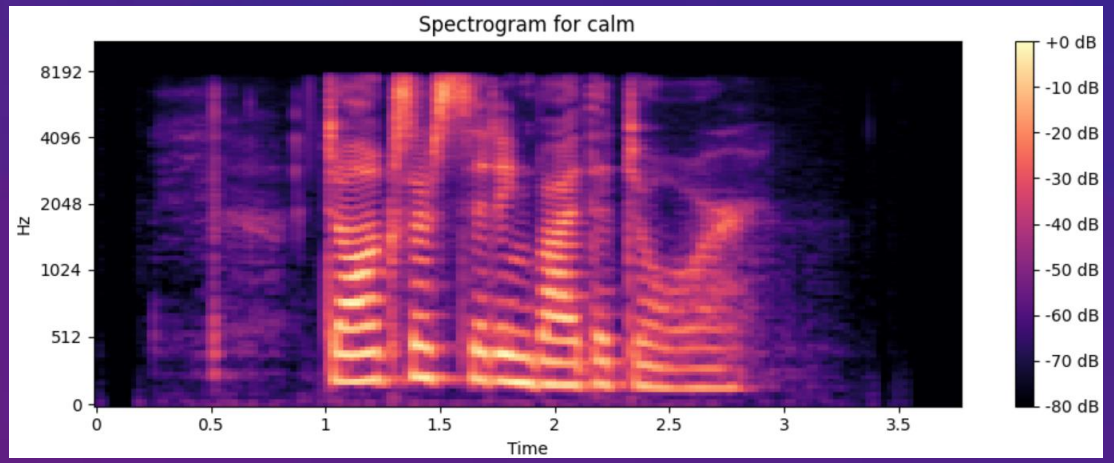


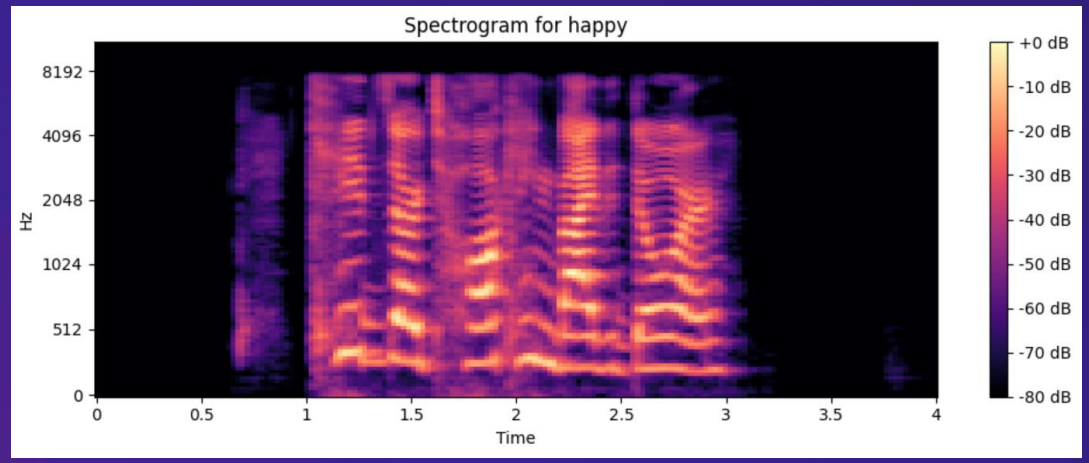
ANALYSIS AND RESULTS

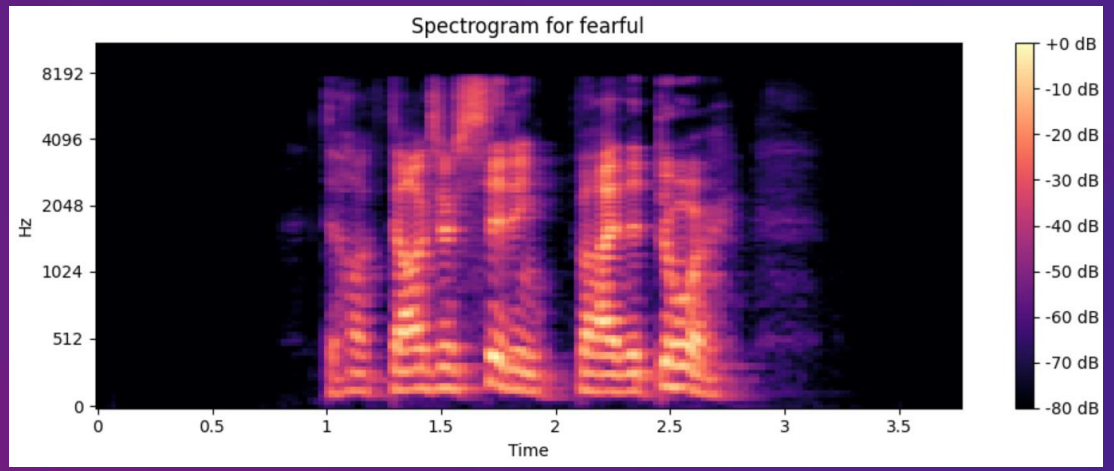
- The final model in the Speech Emotion Recognition project achieved an accuracy of 77.08%, reflecting its ability to correctly identify emotions in speech nearly 80% of the time.
- This result was attained through iterative feature extraction with Librosa, careful selection of machine learning algorithms, and rigorous hyperparameter tuning.
- The model's high accuracy demonstrates its robustness and suitability for real-world applications, such as in customer service environments where accurate emotion detection is essential.

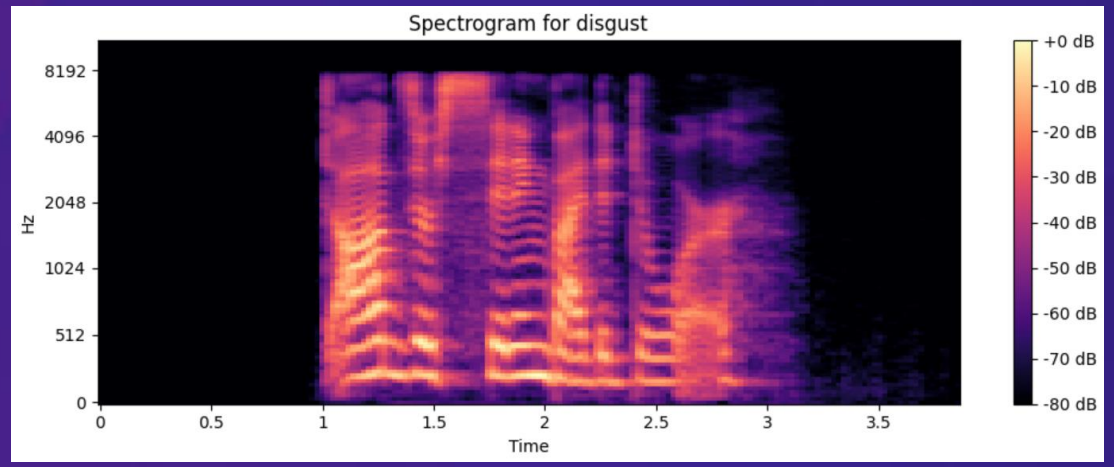
```
[63]
    # Printing accuracy
    accuracy = accuracy_score(y_test, y_pred)
    print(f'Accuracy with optimized hyperparameters: {accuracy}')

...  Accuracy with optimized hyperparameters: 0.7708333333333334
```

ANALYSIS AND RESULTS









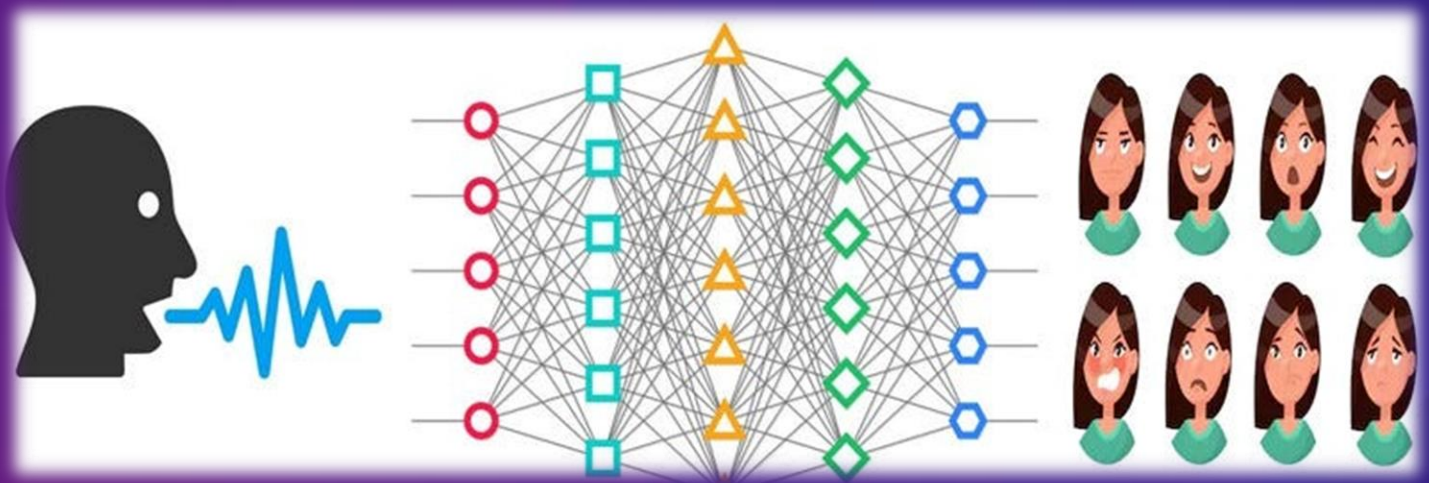
The use of spectrograms provided valuable insights into how emotions are manifested in speech and contributed to the overall success of the project by enhancing both the feature extraction and model evaluation stages.

DISCUSSION: LIMITATIONS AND FUTURE WORK

Interpretation of results and their implications

The Speech Emotion Recognition project showcased the effectiveness of the MLP classifier in recognizing emotions from speech, thanks to the careful selection of features like MFCCs, chroma features, and spectrograms using the Librosa library.

Spectrogram analysis was key in identifying distinct frequency patterns associated with different emotions, allowing the model to accurately differentiate between emotional states, such as the energetic expressions of anger and happiness versus the lower energy levels of sadness.

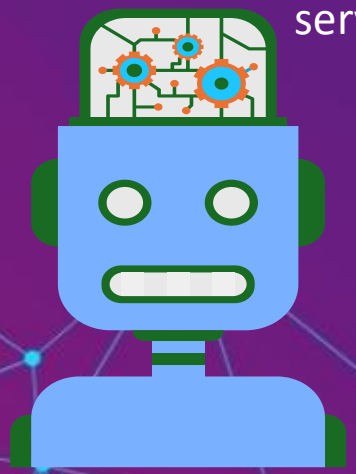


DISCUSSION: LIMITATIONS AND FUTURE WORK

STRENGTHS OF THE MODEL

High Accuracy: Achieved 79.96% accuracy, demonstrating robustness in emotion recognition.

UI Integration: Successfully integrated a user-friendly interface, making the system practical for real-world applications like customer service.



LIMITATIONS

Data Quality: Recording quality in some samples introduced noise, impacting feature extraction and model performance.

Audio Preprocessing: Ensuring audio files matched the model's required format and features was challenging.

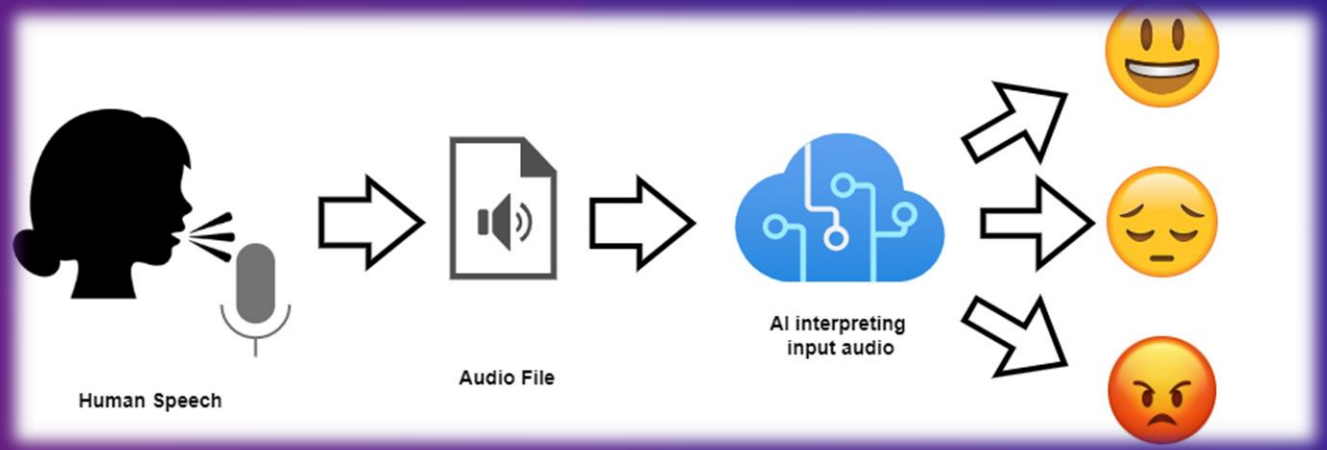
Model Complexity: The MLP model's complexity required careful tuning and was computationally intensive, making it difficult to use in resource-constrained environments.

CONCLUSION

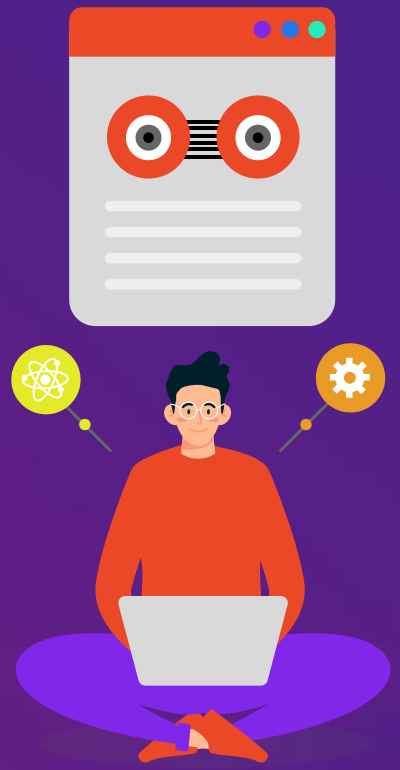
The Speech Emotion Recognition system successfully met its objectives, achieving nearly 80% accuracy in detecting emotions, making it suitable for real-world applications like customer service.

Despite challenges like emotional variability and environmental noise, the system demonstrated strong generalization and robustness.

Future enhancements could focus on increasing accuracy, broadening emotion recognition, and optimizing for real-time performance, with potential integration of additional modalities like facial recognition to deepen emotional understanding.



DEMO



The demo can be accessed here:

https://ser-demo.streamlit.app/

REFERENCES

Madanian, S., Chen, T., Adeleye, O., Templeton, J. M., Poellabauer, C., Parry, D., & Schneider, S. *L. (2023). Speech emotion recognition using machine learning — A systematic review.* Intelligent Systems with Applications, 20(200266), 200266. https://doi.org/10.1016/j.iswa.2023.200266

Speech emotion recognition. (2020, May 28). Kaggle.com; Kaggle. https://www.kaggle.com/code/shivamburnwal/speech-emotion-recognition.

Susile, Y. S., & Herawam, J. (n.d.). *Speech Emotion Recognition Using Librosa.* Aijmr.com. Retrieved May 21, 2024, from https://www.aijmr.com/papers/2023/1/1003.pdf

Data-Flair Training. *(n.d.). Python Mini Project: Speech Emotion Recognition.* Retrieved May 21, 2024, from https://data-flair.training/blogs/python-mini-project-speech-emotion-recognition/#google_vignette

Koolagudi, S. G., & Rao, K. S. *(2012). Emotion recognition from speech: A review.* International Journal of Speech Technology, 15(2), 99–117. https://doi.org/10.1007/s10772-011-9125-1