Speech Emotion Recognition

Niket Agrawal(M23CSA520)
Ritesh Lamba (M23CSA544)

1. Task Explanation and Importance

(a) Task Overview

Speech Emotion Recognition (SER) involves detecting emotions from spoken language, such as happiness, sadness, anger, or neutral tones. The goal is to classify audio samples into predefined emotion categories based on acoustic features extracted from speech signals.

(b) Importance in Real-World Applications

SER has significant applications in various domains:

- Customer Service : Analyzing customer sentiment during calls to improve service quality.
- Mental Health : Detecting emotional distress in therapy sessions or mental health apps.
- Human-Computer Interaction : Enabling AI systems to respond empathetically in virtual assistants or chatbots.
- Security : Identifying stress or aggression in surveillance systems.

2. State-of-the-Art Models and Tools

(a) Traditional Machine Learning Models

- Methods : SVM, Random Forest, k-NN using handcrafted features (e.g., MFCCs, Chroma, Spectral Contrast).
- Strengths :
    - Simple and interpretable.
    - Effective for small datasets.
- Limitations :
    - Limited ability to capture complex patterns in audio data.
    - Requires manual feature engineering.

(b) Deep Learning Models

1. CNNs (Convolutional Neural Networks) :
    - Use Mel-Spectrograms or spectrograms as input.

- Strengths: Captures spatial patterns in spectrograms effectively.
- Limitations: Struggles with temporal dependencies in speech.

2. RNNs/LSTMs (Recurrent Neural Networks/Long Short-Term Memory) :
    - Process sequential data (e.g., MFCCs over time).
    - Strengths: Handles temporal dependencies well.
    - Limitations: Computationally expensive; prone to vanishing gradients.
3. Transformer-Based Models :
    - Examples: Wav2Vec 2.0 fine-tuned for emotion recognition.
    - Strengths: Captures long-range dependencies; highly accurate.
    - Limitations: Requires large datasets and computational resources.

3. Results and Metrics

(a) Metrics Used

- Accuracy : Measures overall correctness of predictions.
- F1-Score : Balances precision and recall, especially useful for imbalanced datasets.
- Confusion Matrix : Highlights misclassifications between emotion classes.

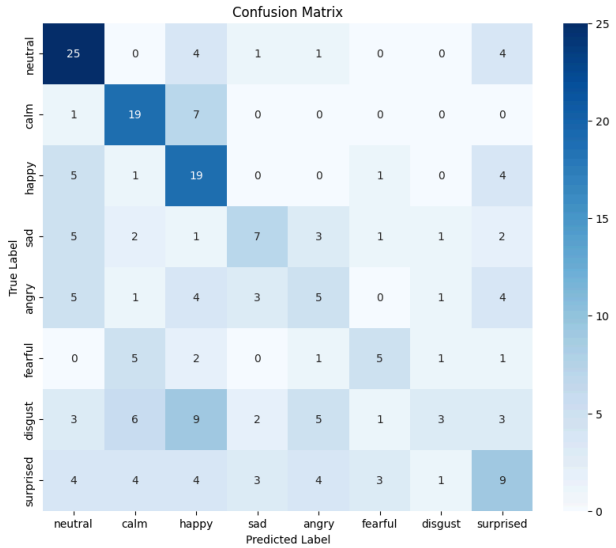
We have used the ravdess dataset and implemented two models one with using only MFCC feature and another with combined MFCC, Chroma, Spectral , Mel Spectrogram.

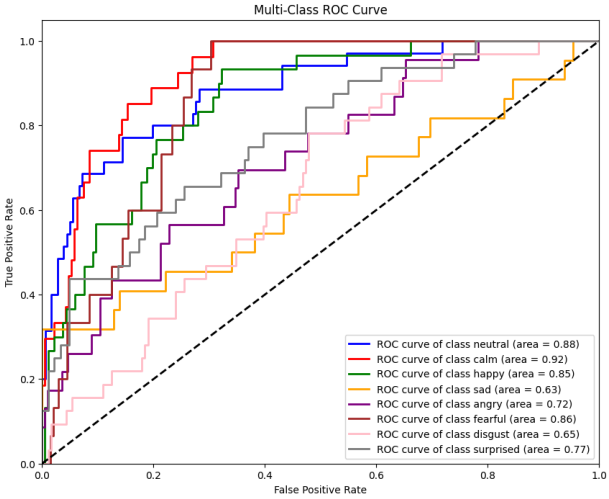
Github Repository link : https://github.com/M23CSA520/Speech_Understanding_PA1
Code : https://github.com/M23CSA520/Speech_Understanding_PA1/blob/main/m23csa520_speech_q1_pa1.py

Results Summary

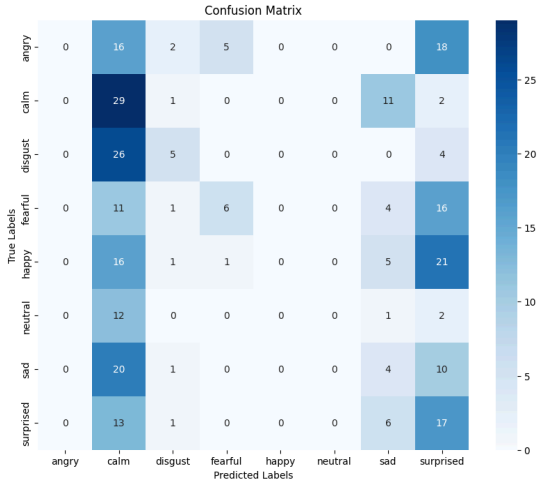
Feature Set	Accuracy %	F1-Score
Only MFCC	42.59	0.3942
Combined	21.18	0.1532

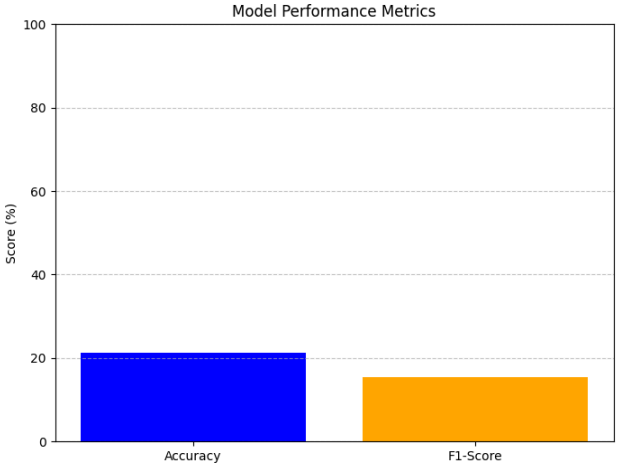
Confusion Matrix & Roc Curve for only MFCC





Confusion Matrix for Combined Features





4. Discussion of Results

(a) Key Observations

1. MFCCs Alone Perform Better :
    - Using only MFCCs achieves 42.59% accuracy and an F1-score of 0.3942 , demonstrating the effectiveness of this feature set.
    - Adding Chroma, Spectral Contrast, and Mel-Spectrogram degrades performance, likely due to:
        - Overfitting to noise introduced by additional features.
        - Misalignment between features during preprocessing.
        - The model architecture does not effectively leverage diverse feature types.
2. Class Imbalance :
    - Some emotions (e.g., "Fearful" and "Disgust") are underrepresented in the dataset, leading to poor performance for these classes.
    - The confusion matrix shows that the model frequently predicts "Calm" when using all features, indicating a bias toward this class.
3. Model Bias :
    - The model tends to predict "Calm" frequently, especially when using all features. This could be due to:
        - Class imbalance.
        - Poor generalization caused by noisy or irrelevant features.

(b) Strengths of Metrics

- Accuracy : Provides a high-level overview of performance.
- F1-Score : Accounts for class imbalance and balances precision and recall.
- Confusion Matrix : Highlights specific strengths and weaknesses for each emotion class.

(c) Limitations of Metrics

- Accuracy : Can be misleading for imbalanced datasets.
- Confusion Matrix : Lacks quantitative aggregation but provides detailed insights.

5. Open Problems and Opportunities

(a) Open Problems

1. Class Imbalance :
    - Some emotions are underrepresented in datasets, leading to poor performance for minority classes.
2. Cross-Cultural Variability :
    - Emotions expressed differently across cultures may reduce model generalizability.
3. Noise Robustness :
    - Models struggle in noisy environments, limiting real-world applicability.

(b) Opportunities

1. Develop Transformer-Based Models :
    - Fine-tune models like Wav2Vec 2.0 for SER to leverage their ability to capture long-range dependencies.
2. Create Larger, More Diverse Datasets :
    - Include more samples for minority classes and diverse cultural expressions.
3. Explore Multimodal Approaches :
    - Combine audio with facial expressions or text for improved emotion recognition.

6. Conclusion

In this project, we implemented a Speech Emotion Recognition (SER) system using features such as MFCCs, Chroma, Spectral Contrast, and Mel-Spectrogram. Our results show that using only MFCCs outperforms combining all features, achieving 42.59% accuracy and an F1-score of 0.3942 . However, challenges such as class imbalance and model bias remain. Future work should focus on addressing these issues through improved datasets, model architectures, and multimodal approaches.