Emotion Detection from Speech Using Machine Learning

Maha Baig
Department of Computing
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
mbaig.bscs22seecs@seecs.edu.pk

Hamza Riaz
Department of Computing
National University of Sciences and Technology (NUST)
Islamabad, Pakistan
hriaz.bscs22seecs@seecs.edu.pk

***Abstract*—Emotion detection from speech bridges the gap between human communication and artificial intelligence, offering non-invasive methods to analyze emotional cues. This study focuses on classical machine learning models, specifically Random Forests and Support Vector Machines (SVMs), to classify emotions such as `angry`, `sad`, `happy`, and `fearful`. Leveraging well-known datasets like RAVDESS, TESS, SAVEE, and CREMA-D, the project emphasizes robust feature engineering, extracting acoustic features like MFCCs, jitter, shimmer, and spectral properties.**

To address data diversity and balance, techniques such as pitch shifting, noise addition, and SMOTE were employed for data augmentation and class balancing. These methods not only enhanced model robustness but also mitigated challenges arising from dataset limitations, including variability in accents and recording conditions.

The models achieved strong accuracy after combining the datasets, demonstrating their effectiveness in capturing distinct emotional features. However, performance declined on some emotions like `fearful` and `disgust`, revealing overlaps in feature spaces for them. Error analysis using PCA and clustering visualizations confirmed these overlaps, suggesting the need for advanced feature extraction techniques or deep learning approaches to improve separability.

The findings highlight scalable applications in mental health monitoring, virtual assistants, and education, where emotion-aware systems can enhance human-computer interaction. Despite its limitations, this study provides a solid foundation for future work, including the integration of deep learning and multimodal approaches to further refine emotion recognition systems.

***Index Terms*—Emotion detection, machine learning, RAVDESS, TESS, SAVEE, CREMA-D, speech analysis**

I. Introduction

A. *Background*

Emotions are fundamental to human communication, shaping decisions, behaviors, and interactions. For human-machine interfaces to be effective, understanding and responding to emotional cues is essential. Speech Emotion Recognition (SER) offers a scalable and non-invasive approach to extracting emotional insights from audio signals, positioning itself as a vital component in affective computing and artificial intelligence. By bridging the emotional gap in human-machine interactions, SER systems enhance natural communication and pave the way for emotion-aware technologies.

B. *Applications*

SER systems have diverse applications:

- **Healthcare**: Diagnosing mental health conditions and monitoring patient well-being.
- **Education**: Detecting student engagement and stress levels in real-time.
- **Customer Service**: Enhancing interactions in call centers by recognizing customer satisfaction or frustration.
- **Human-Machine Interaction**: Enabling emotion-aware virtual assistants and social robots.

C. *Objectives*

The goal of this project was to develop a robust, scalable SER system using classical machine learning algorithms. By leveraging audio features and addressing dataset limitations, the project aims to improve emotion detection accuracy in resource-constrained environments.

II. Literature Review

A. *Background and Significance*

Emotion recognition, a key focus of affective computing, aims to enable machines to understand and respond to human emotions. Early work by Ververidis and Kotropoulos (2006) highlighted the importance of acoustic features like pitch, MFCCs, and speech energy, forming the foundation of traditional machine learning models.

Recent advancements focus on deep learning, such as Sadok et al. (2023)'s Vector Quantized Masked Autoencoder (VQ-MAE-S), which leverages large datasets like VoxCeleb2 for self-supervised learning. While effective, these methods require substantial computational resources and annotated datasets, limiting their feasibility in real-time or low-resource settings.

B. *Classical vs Deep Learning Approaches*

Classical machine learning models like SVMs, Random Forests (RFs), and k-Nearest Neighbors (KNNs) provide lightweight alternatives to deep learning, excelling with robust feature engineering. SVMs achieve high accuracy in high-dimensional spaces, while RFs handle feature variability and noise effectively.

In contrast, deep learning has revolutionized emotion recognition with end-to-end learning from raw audio. Models like CNNs and RNNs deliver state-of-the-art performance but require large datasets and significant computational resources, limiting their use in resource-constrained environments.

C. Feature Engineering and Data Augmentation

With limited data, feature extraction and augmentation were crucial. MFCCs, spectral features, and vocal quality metrics (e.g., jitter, shimmer) effectively captured emotional nuances. Augmentation techniques like pitch shifting and noise addition addressed dataset limitations and enhanced model robustness.

D. Challenges and Open Questions

Key challenges in emotion recognition include:

- **Dataset Limitations**: Small dataset sizes and limited diversity in accents and emotional expressions.
- **Generalization**: Models often struggle to perform well across datasets due to variations in recording conditions and speaker characteristics.
- **Interpretability**: Classical models offer more transparency in feature importance compared to deep learning models, which are often criticized as ”black boxes.”

E. Relevance to This Project

This project aligns with the current trend of exploring lightweight, scalable models for real-world applications. By emphasizing feature engineering and augmentation, it addresses the limitations of small datasets and aims to develop models that generalize well across diverse datasets. Unlike multimodal approaches that integrate audio and visual data, this project focuses solely on audio features, making it suitable for scenarios where visual data is unavailable.

III. Dataset Overview

Domain Knowledge Acquisition

To build foundational knowledge in audio processing, we undertook a comprehensive 10-hour course by Valerio Velardo on YouTube. This course delved into the theory behind essential audio features such as MFCCs, spectral features, pitch, jitter, and shimmer. It emphasized their significance in capturing human-perceived emotional nuances from speech. Key topics covered included:

Feature Theory and Usage: Understanding the purpose of features like spectral centroid and MFCCs in emotion recognition.

Mathematical Foundations: Exploring concepts such as the Discrete Fourier Transform (DFT) to analyze frequency components in audio signals.

Correlation with Human Perception: Studying how features like pitch and intensity mirror human auditory perception.

Feature Extraction Techniques: Learning methodologies to extract and interpret features using tools such as Librosa.

This foundational knowledge equipped us with a robust understanding of audio processing and its application to speech emotion recognition, allowing us to design feature engineering and preprocessing steps effectively.

In our problem, four datasets were utilized:

- **RAVDESS**: Comprising 1,440 audio files from 24 actors, including male and female speakers. The dataset predominantly features speakers of North American descent, offering a controlled recording environment with minimal background noise. File names are structured to encode metadata such as emotion, intensity, and actor ID, e.g., ”03-01-05-02-02-01-12.wav,” where each segment represents specific attributes of the recording. The dataset includes 8 distinct emotion classes: neutral, calm, happy, sad, angry, fearful, disgust, and surprised.
- **SAVEE**: Containing 480 recordings from 4 male actors, all of British ethnicity, this dataset emphasizes clear articulation of emotions. File names follow a pattern reflecting emotion categories and actor identifiers, simplifying data organization. The dataset comprises 7 emotion classes: neutral, happy, sad, angry, fearful, disgust, and surprised.
- **CREMA-D**: Featuring 7,442 audio clips from 91 actors of diverse ethnic backgrounds, CREMA-D incorporates variability in accents and vocal styles. File names encode speaker ID, emotion, and other attributes, aiding in detailed analysis and dataset management. This dataset includes 6 emotion classes: neutral, happy, sad, angry, fearful, and disgust.
- **TESS**: Offering 5,600 emotional speech samples from 2 actresses, both of Canadian descent, TESS is characterized by high-quality recordings. The files are named to reflect emotion and prompt, e.g., `"Angry_Dog_01.wav"` ensuring intuitive categorization. The dataset includes 7 emotion classes: neutral, happy, sad, angry, fearful, disgust, and surprised.

The combined dataset includes **15,942 samples**, ensuring diverse data sources for robust model training.

For consistency we chose the following **7 classes** for classification in all our testing: **neutral, happy, sad, angry, fearful, disgust, and surprised**.

IV. Dataset Exploration

Before we even implemented a single model, we explored our dataset comprehensively. We wanted to see the distribution of our data and whether we are able identify any patterns or trends that would suggest where the model might perform well or struggle. Here are our key findings:

A. Correlation Matrix Analysis

The correlation matrix revealed critical relationships among the features in the dataset:

1) **High Correlation Among MFCCs:**
    - Many MFCC features (e.g., `MFCC_2`, `MFCC_4`, `MFCC_6`) exhibited high inter-correlations, including their derivatives (`Delta_MFCC`, `Delta2_MFCC`). This redundancy suggests that dimensionality reduction techniques like

Principal Component Analysis (PCA) could effectively capture the most meaningful variations while reducing noise.

2) **Spectral Features:**
   - Features like `Spectral_Flatness` and `Spectral_Contrast` demonstrated low-to-moderate correlations with MFCCs and chroma features. These unique features provide complementary information that may enhance classification accuracy.
3) **Distinct Feature Clusters:**
   - Chroma features formed a moderately correlated cluster, distinct from MFCCs, indicating their potential as valuable pitch-related predictors.
4) **Low Correlation with Duration:**
   - The `Duration` feature showed low correlations with other features, suggesting its independence and potential as a unique predictor in distinguishing emotions.
5) **Jitter and Shimmer:**
   - Both `Jitter` and `Shimmer` were weakly correlated with MFCCs and chroma features, reflecting their unique ability to capture voice quality aspects such as pitch irregularities and amplitude variations.

B. Feature Distribution Insights

1) **Root Mean Square Energy (RMSE):**
   - Observation: Highly right-skewed with most values concentrated at lower ranges.
   - Insight: Indicates the majority of audio signals represent low-energy emotions (e.g., `neutral`, `calm`). RMSE effectively separates high-energy emotions like `angry` and `happy` from low-energy ones like `sad`.
2) **Spectral Features:**
   - Spectral Contrast: Shows a roughly normal distribution peaking around 20 and 3000Hz, useful for distinguishing high-intensity emotions (`angry`) from subdued ones (`neutral`).
   - Spectral Flatness: Right-skewed, indicating most samples are tonal, aligning with emotions like `happiness` compared to noisier ones like `anger`
3) **Chroma Features:**
   - Nearly normal distributions for `Chroma_1` and `Chroma_4` with slight variations, reflecting pitch-related characteristics critical to emotions like happiness and fear.
4) **MFCCs:**
   - MFCC_1: Symmetrical distribution, centered, indicative of timbral properties.
   - MFCC_2: Multi-modal, suggesting clusters potentially linked to specific emotions.

C. Class-Wise Feature Analysis

1) **RMSE:**
   - High median values for `angry` and `happy` versus lower values for `sad` and `neutral` highlight its role in capturing emotional intensity.
2) **Spectral Contrast:**
   - Slightly elevated for `angry` and `happy`, reflecting variability in vocal intensity.
3) **Jitter and Shimmer:**
   - `Angry` and `sad` have higher jitter, suggesting stressed or irregular vocalizations, whereas `happy` exhibits higher shimmer, capturing amplitude variability
4) **MFCCs:**
   - MFCC_1 effectively separates broad emotional categories like `angry` and `happy`, while MFCC_2 uniquely distinguishes surprised due to its compact distribution.
5) **Chroma Features:**
   - Higher median values for `happy` and `surprised` suggest their association with tonal pitch variations.

D. PCA Analysis

1) **Explained Variance:**
   - The first principal component (PC1) explains approximately 20% of the variance, while PC2 captures slightly less. Together, PC1 and PC2 account for significant variability, with diminishing returns beyond the first few components.
2) **Scatter Plot (PC1 vs. PC2):**
   - Distinct clusters are visible for emotions like `angry` and `neutral`, while overlapping regions for `happy` and `surprised` highlight shared acoustic properties.
   - PC1 primarily separates intensity-driven emotions, and PC2 adds tonal differentiation.

E. Feature Importance Analysis

1) **Key Features:**
   - `RMSE`, `HNR`, `MFCC_1`, and `MFCC_2` were identified as the most discriminative, capturing intensity, voice quality, and timbral distinctions effectively.
2) **Chroma Features:**
   - Contributed to pitch-related emotional nuances, particularly for `happy` and `surprised`.
3) **Shimmer and Jitter:**
   - Captured subtle variations in voice quality, aiding in the differentiation of emotions with expressive vocalizations like `happy` and subdued ones like `neutral`.

F. Insights from Outlier Detection

1) **Emotion-Wise Outliers:**
   - `Angry` and `sad`exhibited the highest outlier counts, reflecting the diversity in their acoustic expressions.
   - `Surprised` had the least outliers, indicating more consistent feature patterns.
2) **Feature-Specific Observations:**
   - RMSE outliers indicated high-energy extremes, particularly for `angry` and `happy`.
   - Jitter and shimmer variability suggested potential preprocessing noise or authentic expressive differences.

G. Temporal Feature Trends

1) **RMSE Trends:**
   - `Angry` showed extreme peaks, representing fluctuating energy, while `happy` exhibited smoother variations..
2) **MFCC_1 and MFCC_2:**
   - Distinct patterns for `angry` and `happy`, with separation in tonal and timbral properties, indicating these features' utility in emotion classification.

H. Key Takeaways

1) **Highly Discriminative Features:**
   - `RMSE`, `HNR`, `MFCC_1`, and `MFCC_2` were pivotal for distinguishing emotions.
2) **PCA and Redundancy Insights:**
   - PCA confirmed the dominance of key components, while redundancy analysis suggested potential feature pruning for improved efficiency.
3) **Challenges:**
   - Overlaps in features like `Chroma` and `MFCC_2` highlighted challenges in distinguishing certain emotions (`fearful`, `disgust`).

By combining these insights, the dataset exploration provided us with a strong foundation for understanding our dataset.

V. Methodology

A. Feature Extraction

A total of **67 features** were extracted from each of the 4 datasets. Key features were extracted using the **Librosa** library, focusing on:

Audio Signal Features

- **Root Mean Square Energy (RMSE)**: Represents signal intensity.
- **Zero Crossing Rate**: Measures the noisiness of the signal.
- **Mel-Spectrogram Mean**: Highlights the frequency content of audio.

MFCC Features:

- **MFCCs (1-13)**: Captures timbral properties essential for speech analysis.
- **Delta and Delta-Delta Coefficients**: Represents temporal changes.

Spectral Features:

- **Spectral Centroid**: Indicates brightness of the sound.
- **Spectral Contrast, Bandwidth, Flatness**: Capture tonal and noise-like properties.
- **Mel-Spectrogram Mean**: Highlights the frequency content of audio.

Chroma Features:

- **Chroma 1-12**: Represents energy across the 12 pitch classes.

Pitch and Vocal Features:

- **Pitch Mean, Harmonic-to-Noise Ratio (HNR), Jitter, Shimmer**: Analyze vocal periodicity and amplitude variations.

Note: Audio Duration was kept constant among all 15,942 samples to prevent the model from learning to use it as a sole differentiating factor among emotions.

B. Data Preprocessing

Emotion Mapping: Emotions across datasets were mapped to a consistent set of labels (e.g., "happy," "sad," "angry") based on file names to standardize classification targets. This ensured all emotions were uniformly labeled without altering their original semantics. No mapping was changed to a different emotion. We just extracted all the emotions that were available, without adding any new label on our own.

Augmentation: Techniques such as noise addition, pitch shifting, and time stretching were employed to mitigate dataset size limitations and enhance model robustness

Splitting: A stratified 5-fold cross-validation strategy ensured equal representation of all classes in training and test sets.

Normalization: Standardized feature values to ensure consistency across datasets.

Class Balancing: Synthetic Minority Over-sampling Technique (SMOTE) addressed class imbalance by generating synthetic samples for underrepresented classes.

C. Models and Evaluation

Classical machine learning algorithms, including Support Vector Machine (SVM) and Random Forest (RF), were implemented. Hyperparameter tuning via GridSearchCV optimized the models, which were then evaluated using metrics such as accuracy, precision, recall, and F1-score. The confusion matrix was also used to see how each of the 7 classes were being classified by the model.

We conducted a total of five experiments to comprehensively evaluate model performance. The first three experiments focused on running the same pipeline on individual datasets, allowing us to assess how the models performed when trained and tested on a single dataset in isolation.

In the fourth experiment, we combined all datasets into a single, larger dataset to train a generalized model, aiming to understand its ability to handle diverse data sources and contexts. Finally, the fifth experiment explored the impact of data augmentation techniques on model performance, examining whether these techniques could enhance the model's robustness and accuracy.

VI. Results

Initially, our aim was to create a model trained on the RAVDESS dataset only. However, upon finding the other datasets, we decided to experiment with them too. The same pipeline of feature extraction and pre-processing was applied on all the datasets individually. Both Random Forests and Support Vector Machines (SVM) were trained and fine-tuned on them as well.

A. Individual Datasets Results

Model	**RAVDESS**	**SAVEE**	**TESS**	**CREMA-D**
Random Forest	92%	83.0%	99.0%	49.0%
SVM	94%	74.0%	99.0%	50.0%

TABLE I
Accuracies of fine-tuned models on different datasets

1) Random Forest and SVM models showed strong performance on datasets with controlled recordings and clear emotional expressions, achieving accuracies above 90% in some cases.
2) Variability in accents, recording conditions, and smaller dataset sizes led to significant performance drops, particularly on datasets with more diverse or challenging characteristics.
3) Both models excelled in handling well-structured data but struggled to generalize across datasets with higher variability, highlighting the need for improved feature engineering and augmentation techniques.

B. Combined Datasets Results

1) **Baseline Models:**
   - **Random Forest:**
     - Tuned Random Forest results are presented in Table II.
     - The confusion matrix for the base Random Forest model is shown in Figure 1.
     - **Analysis:**
       * Random Forest achieved high precision for `Surprised` and `Angry`, indicating its ability to handle distinct emotional cues effectively.
       * Moderate performance on `Happy` and very poor performance on `Fearful` suggests challenges in differentiating between these emotions.
       * The overall weighted accuracy of 72% highlights the model's reasonable generalization across the four datasets.
   - **Support Vector Machine (SVM):**
     - Optimized SVM model performance results are summarized in Table III.

TABLE II
Tuned Random Forest Results

Class (Emotion)	**Precision**	**Recall**	**F1-Score**	**Support**
0 (`Angry`)	0.80	0.84	0.82	491
1 (`Disgust`)	0.72	0.64	0.68	491
2 (`Fearful`)	0.53	0.44	0.48	331
3 (`Happy`)	0.70	0.74	0.72	491
4 (`Neutral`)	0.71	0.73	0.72	493
5 (`Sad`)	0.76	0.81	0.78	503
6 (`Surprised`)	0.74	0.87	0.80	89
Accuracy	0.72 (Weighted Average)			

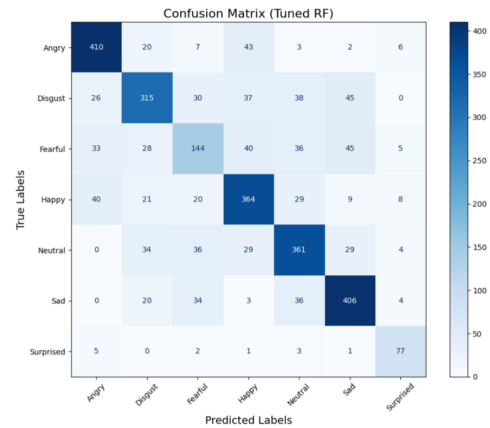

Fig. 1. Confusion Matrix for the Base Random Forest Model.

     - The confusion matrix for the SVM model is shown in Figure 2.
     - **Analysis:**
       * SVM performed well for `Surprised` and `Angry`, reflecting its strength in separating distinct classes in a high-dimensional space.
       * Lower scores for `Disgust` and `Fearful` suggest that model is unable to learn the very minute emotional cues distinguishing these emotions.
       * An overall accuracy of 70% indicates good generalization but slightly lags behind Random Forest in weighted average metrics.

TABLE III
Optimized SVM Model Performance

Class (Emotion)	**Precision**	**Recall**	**F1-Score**	**Support**
0 (`Angry`)	0.81	0.80	0.81	491
1 (`Disgust`)	0.57	0.72	0.63	491
2 (`Fearful`)	0.56	0.43	0.48	331
3 (`Happy`)	0.73	0.70	0.72	491
4 (`Neutral`)	0.76	0.69	0.72	493
5 (`Sad`)	0.71	0.73	0.72	491
6 (`Surprised`)	1.00	0.90	0.95	77
Accuracy	0.70 (Weighted Average)			

2) **After Data Augmentation:**
   - **Random Forest:**
     - Results for Random Forest after data augmentation are presented in Table IV.

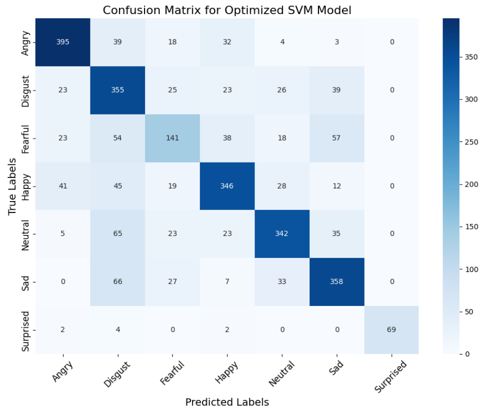

Fig. 2. Confusion Matrix for the Base SVM Model.

- The confusion matrix for Random Forest after data augmentation is shown in Figure 3.
- **Analysis:**
  - ∗ Data augmentation improved recall for `Fearful` but then caused issues with `Happy`.
  - ∗ Precision and F1-scores for `Surprised` and `Sad` remained consistent, reinforcing the model's ability to capture these emotions well.
  - ∗ The overall accuracy was 71%, demonstrating good generalization across the datasets.

TABLE IV
RANDOM FOREST RESULTS AFTER DATA AUGMENTATION

Class (Emotion)	Precision	Recall	F1-Score	Support
0 (`Angry`)	0.77	0.83	0.80	524
1 (`Disgust`)	0.52	0.97	0.67	31
2 (`Fearful`)	0.72	0.66	0.69	524
3 (`Happy`)	0.53	0.42	0.47	363
4 (`Neutral`)	0.71	0.66	0.69	522
5 (`Sad`)	0.70	0.74	0.72	508
6 (`Surprised`)	0.74	0.75	0.74	565
Accuracy	0.71 (Weighted Average)			

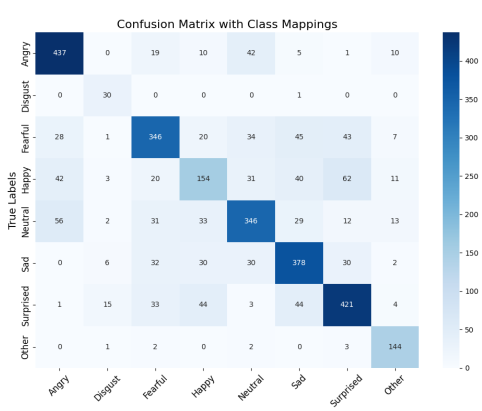

Fig. 3. Confusion Matrix for Random Forest After Data Augmentation.

- **Support Vector Machine (SVM):**
  - Results for SVM after data augmentation are presented in Table V.
  - The confusion matrix for SVM after data augmentation is shown in Figure 4.
  - **Analysis:**
    - ∗ Data augmentation marginally improved performance for `Fearful` but decreased the performance on many of the other classes.
    - ∗ Precision and recall for `Disgust` remained low, signaling potential limitations in SVM for imbalanced datasets.
    - ∗ Overall weighted accuracy stayed at 70%, suggesting minor gains in class-level performance.

TABLE V
SVM RESULTS AFTER DATA AUGMENTATION

Class (Emotion)	Precision	Recall	F1-Score	Support
0 (`Angry`)	0.76	0.81	0.79	524
1 (`Disgust`)	0.58	0.58	0.58	31
2 (`Fearful`)	0.64	0.66	0.65	524
3 (`Happy`)	0.51	0.52	0.52	363
4 (`Neutral`)	0.74	0.67	0.70	522
5 (`Sad`)	0.70	0.73	0.72	508
6 (`Surprised`)	0.75	0.70	0.73	565
Accuracy	0.70 (Weighted Average)			

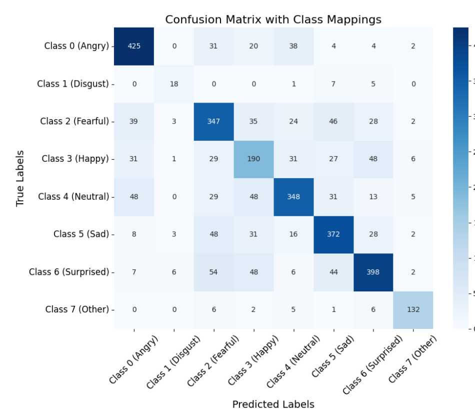

Fig. 4. Confusion Matrix for SVM After Data Augmentation.

C. Error Analysis

To further understand the performance of the models, an error analysis was conducted using two visualizations:

1) **Elbow Method for Optimal Clusters (Figure 5)**: This graph identifies the optimal number of clusters for K-means clustering based on the within-cluster sum of squares (WCSS). The graph indicates that around 5-6 clusters may represent the dataset well. This indicates that there is some overlap in a few classes.
2) **K-means Clustering with PCA (Figure 6)**: The PCA-reduced 2D visualization highlights how the emotions are separated across clusters. However, overlaps between certain clusters, particularly those associated with

Fearful and Disgust, are evident. These overlaps suggest that the feature space for these emotions lacks sufficient separation, most likely contributing to the model's lower performance for these classes.

Observations:

- Fearful and Disgust emotions show significant overlaps in the PCA visualization. This overlap corresponds to their lower precision and recall in the classification results, indicating that the features extracted for these emotions may not be sufficiently distinctive.
- Emotions such as Angry and Surprised exhibit better-defined clusters, reflecting their higher precision and recall values.
- The clustering analysis reinforces the need for advanced feature extraction techniques or additional data preprocessing to improve the separation of Fearful and Disgust in the feature space.

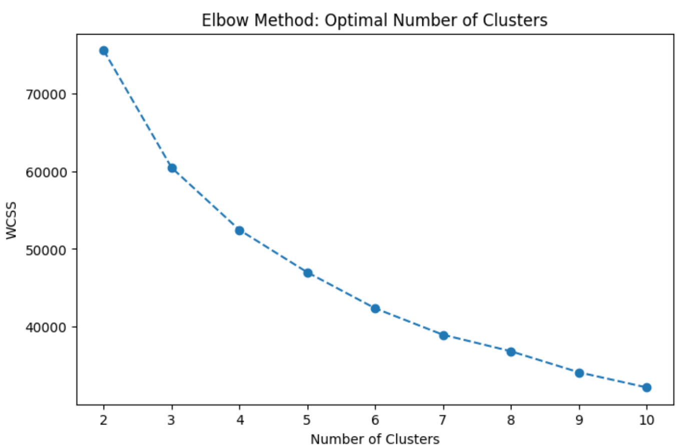

Fig. 5. Elbow Method: Optimal Number of Clusters.

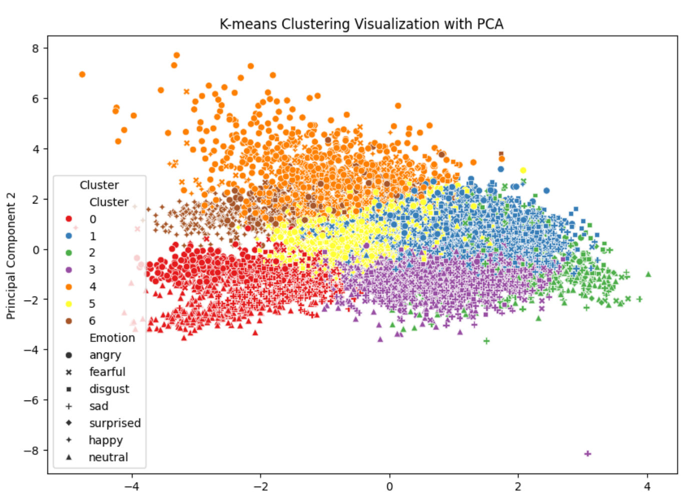

Fig. 6. K-means Clustering Visualization with PCA.

VII. Conclusion

This project demonstrated the viability of classical machine learning models, such as Random Forests and SVMs, for emotion detection from speech. Despite achieving reasonable generalization across datasets, challenges with overlapping emotions like Fearful and Disgust highlight areas for improvement. These results provide a foundation for future work, including advanced feature extraction and deep learning approaches to enhance performance.

References

[1] Ververidis, D., & Kotropoulos, C. (2006). Emotional Speech Recognition: Resources, Features, and Methods.

[2] Sadok, S., Leglaive, S., & Séguier, R. (2023). A Vector Quantized Masked Autoencoder for Speech Emotion Recognition.

[3] George, S. M., & Ilyas, P. M. (2024). A review on speech emotion recognition: A survey, recent advances, challenges, and the influence of noise. *Neurocomputing*, 568, Article 127015. Elsevier.

[4] Valerio Velardo - The Sound of AI (2020). Audio Signal Processing for Machine Learning (Youtube)