Emotion Recognition

Arashdeep Mehroke Mohamed Hasan

Dec 5, 2024

1 Problem Statement

Recognizing human emotions is an essential aspect of understanding and improving human-computer interaction, providing insights into behavioral analysis, mental health monitoring, and customer experience optimization. The **Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)** offers a rich collection of labeled audio data that can be leveraged to develop robust machine learning models for emotion classification.

Despite the availability of advanced feature extraction techniques such as YAMNet embeddings and Librosa-based handcrafted features, achieving high accuracy in emotion recognition remains a challenge due to:

1. **High Dimensionality of Features:** Audio features like YAMNet embeddings are inherently high-dimensional, complicating the training process for traditional models.
2. **Imbalanced Datasets:** Emotional classes often exhibit uneven distributions, leading to biased model performance.
3. **Complexity of Audio Signals:** Variations in pitch, tone, and intensity across speakers make it difficult to accurately classify emotions.
4. **Computational Overheads:** Extracting, visualizing, and analyzing audio features for large datasets requires efficient preprocessing and dimensionality reduction.

This project aims to address these challenges by:

1. Employing state-of-the-art feature extraction techniques (e.g., YAMNet, Librosa).
2. Balancing the dataset using synthetic techniques like SMOTE.
3. Experimenting with various machine learning and deep learning models to identify optimal solutions.
4. Visualizing high-dimensional feature spaces with dimensionality reduction tools like t-SNE and UMAP for better interpretability.

The goal is to build a robust, scalable system capable of accurately classifying human emotions across multiple categories while ensuring computational efficiency and practical applicability.

2 Data Source

The dataset used in this project is the **Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS)**. It consists of 1440 audio files (16-bit, 48kHz `.wav`) featuring 24 professional actors (12 female, 12 male) vocalizing two statements in a neutral North American accent.

2.1 Emotions and Features

The dataset includes 8 emotions: neutral, calm, happy, sad, angry, fearful, surprised, and disgusted, with normal and strong intensities (except neutral). Each statement is repeated twice, resulting in 60 trials per actor.

2.2 Filename Structure

Files are named using a 7-part numerical identifier (e.g., `03-01-06-01-02-01-12.wav`) indicating:

- **Modality:** Audio-only (03).
- **Vocal Channel:** Speech (01).
- **Emotion:** Fearful (06).
- **Intensity:** Normal (01).
- **Statement:** Dogs (02).
- **Repetition:** First (01).
- **Actor:** Actor 12 (female).

3 Methodology

3.1 Data Preprocessing

This study uses audio datasets derived from the RAVDESS dataset, which include features extracted using Librosa and YAMNet libraries. Below is a summary of the datasets used:

- **librosa_balanced.csv:** 45 features, 7119 samples.
- **librosa_extracted_features.csv:** 45 features, 6439 samples.
- **yamnet_balanced.csv:** 1025 features, 7119 samples.
- **yamnet_extracted_features.csv:** 1025 features, 6440 samples.

Feature Exploration

Figure 1 illustrates the distribution of labels in the dataset. Most emotion classes are balanced, but the classes "Neutral" and "Surprise" are underrepresented. For this reason, Synthetic Minority Oversampling Technique (SMOTE) is used to address class imbalances.

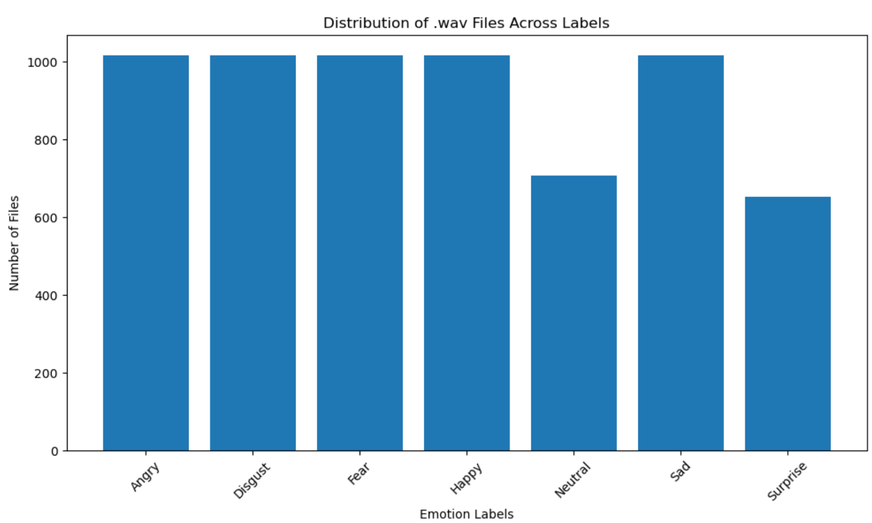

Figure 1: Distribution of emotion labels across audio samples.

Feature distributions across the dataset were analyzed. Figure 2 illustrates the distributions of key extracted features. The distributions reveal a variety of patterns in audio characteristics, which are useful for differentiating emotions.

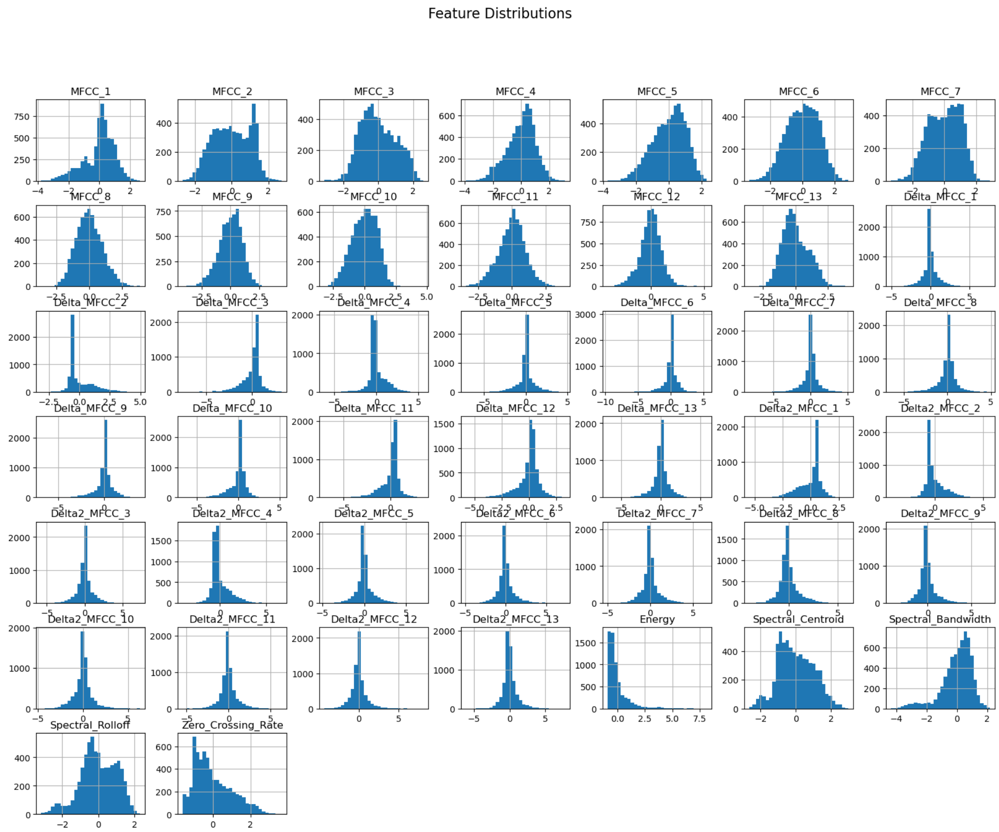

Figure 2: Feature distributions for extracted audio features, including MFCCs, spectral centroid, zero-crossing rate, and energy.

Feature Analysis

Figures 3 and 4 provide insights into the feature distributions and variability for the extracted features.

- **MFCCs:** The Mel-Frequency Cepstral Coefficients (MFCCs) capture timbral information. Peaks and valleys in the MFCC values are indicative of variations in speech timbre across emotions.
- **Spectral Centroid:** Represents the "brightness" of the audio. High centroid values suggest a higher frequency content, while lower values are linked to softer tones.
- **Zero-Crossing Rate (ZCR):** Indicates the noisiness or percussiveness of the signal. Higher ZCR values correspond to noisier audio signals, such as the "Surprise" class.
- **Energy:** Captures the loudness or intensity of the audio signal. Emotions like "Angry" and "Happy" show higher energy levels compared to "Sad" and "Neutral."

Dimensionality Reduction

Dimensionality reduction techniques such as PCA and t-SNE were applied to visualize the feature space. Figure 5 demonstrates clustering of the dataset using PCA. Despite overlapping clusters, PCA highlights trends in feature separability.

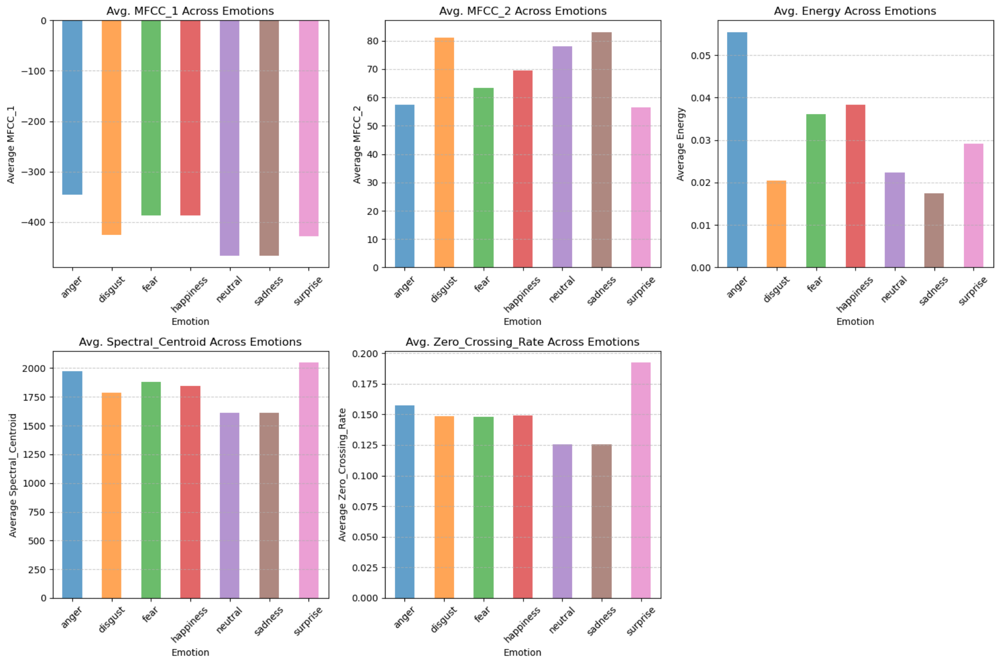

Figure 3: Mean values of MFCCs, Spectral Centroid, Zero-Crossing Rate, and Energy across emotions.

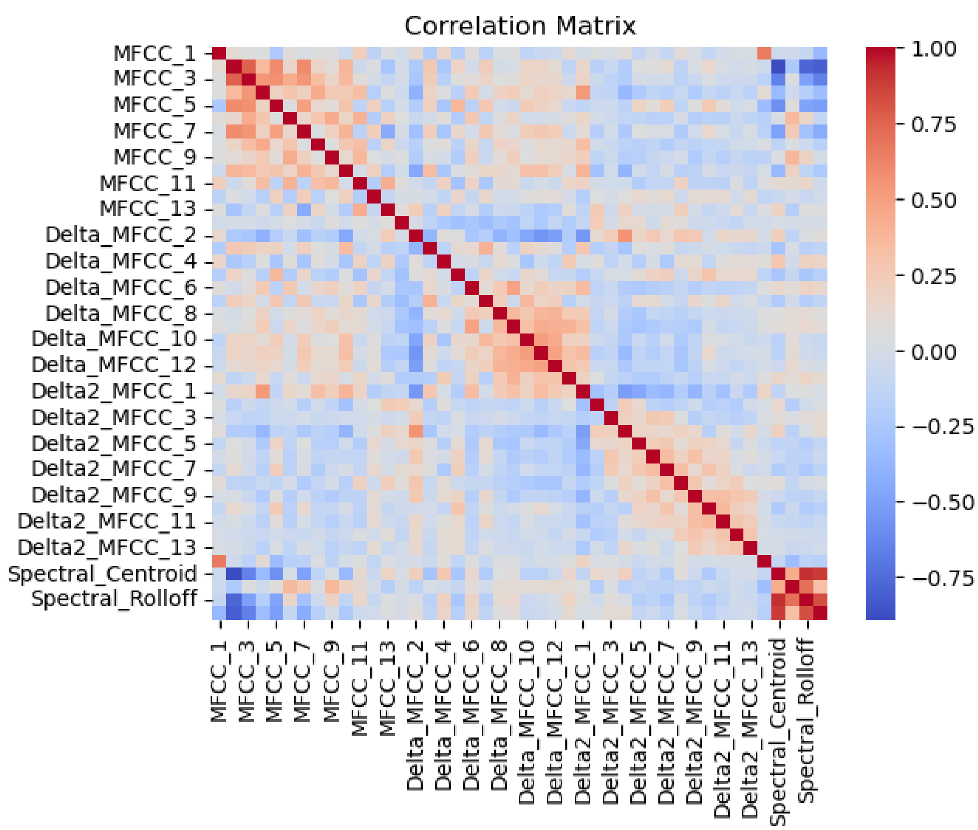

Figure 4: Distribution of feature values across all extracted features.

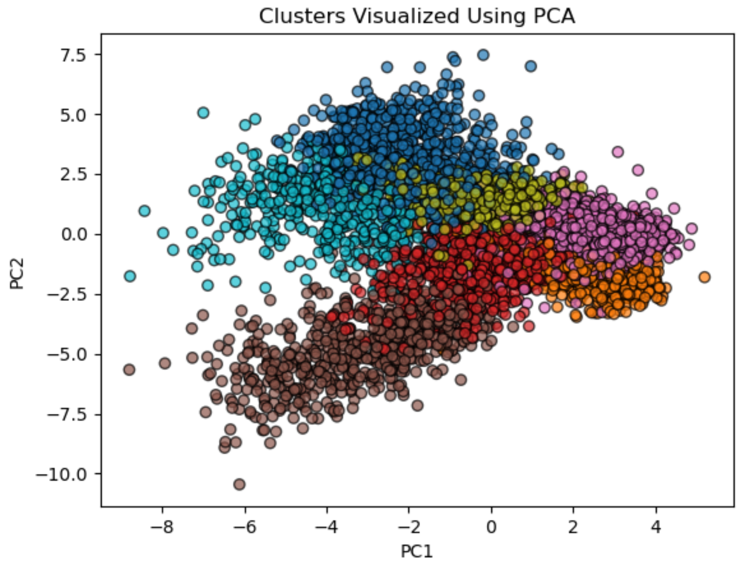

Figure 5: Clusters visualized using PCA. The clustering indicates some separation between emotion classes, though significant overlap is observed.

To address class imbalances, SMOTE (Synthetic Minority Oversampling Technique) was applied, and t-SNE was used to assess the improved separability of the dataset. Figure 6 highlights how t-SNE captures the non-linear relationships among features, improving the clustering of emotions.

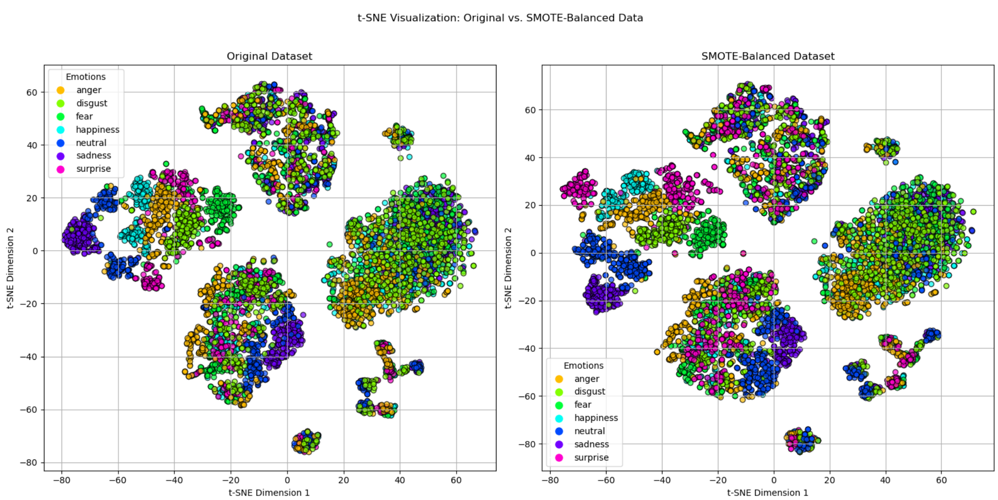

Figure 6: t-SNE visualization of original and SMOTE-balanced datasets. Non-linear feature relationships are effectively captured, enabling better emotion classification.

Model Selection

To classify emotions from the extracted audio features, four machine learning models were employed: K-Nearest Neighbors (KNN), Support Vector Machine (SVC), Random Forest, and a Multi-Layer Perceptron (MLP). Grid search was used to optimize the hyperparameters for KNN, SVC, and Random Forest, while MLP was manually tuned to align with the complexity of the problem.

- **K-Nearest Neighbors (KNN):** KNN was chosen for its simplicity and interpretability. It performs well in scenarios where emotion classes have non-linear boundaries, as it classifies based on the proximity of similar feature patterns. KNN serves as an effective baseline for understanding how well the extracted features group emotions in the feature space.

- **Support Vector Machine (SVC):** SVC is particularly suited for high-dimensional datasets like YAMNet features, as it seeks to find the optimal margin between classes. Its ability to use kernel functions allows it to capture both linear and non-linear decision boundaries, making it a strong candidate for emotion classification where subtle differences in feature values distinguish emotions such as "Happy" and "Fearful."
- **Random Forest:** Random Forest was selected for its robustness and ability to handle heterogeneous data, such as a mix of spectral and temporal features. It combines decision trees to handle complex interactions among features, which is crucial in capturing the intricate variations in speech and audio characteristics associated with different emotions.
- **Multi-Layer Perceptron (MLP):** MLP was chosen for its ability to model complex, non-linear relationships in the data. Emotions often involve subtle, overlapping patterns in features like energy, spectral centroid, and MFCCs, which MLP can effectively learn. Its hierarchical feature learning capability makes it particularly well-suited for tasks like emotion recognition, though it requires careful tuning to prevent overfitting.

Hyperparameter Tuning

Grid search was applied to systematically explore the hyperparameter space for KNN, SVC, and Random Forest. For KNN, the number of neighbors and distance metrics were tuned; for SVC, the kernel type, regularization parameter (C), and kernel coefficient (*gamma*) were optimized; and for Random Forest, the number of trees (*n_estimators*), maximum tree depth (*max_depth*), and minimum samples per split (*min_samples_split*) were adjusted. MLP was tuned manually based on its learning rate, number of hidden layers, and neurons per layer.

Model Evaluation

All models were evaluated using metrics relevant to the emotion classification task: accuracy, precision, recall, and F1-score. These metrics were chosen to balance the trade-offs between false positives and false negatives, which are critical for applications like affective computing. Cross-validation was employed to ensure robust evaluation, particularly given the imbalanced nature of certain emotion classes. The results of these evaluations are discussed in detail in the subsequent sections.

Evaluation and Results

Citation

Livingstone SR, Russo FA (2018). The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS). *PLoS ONE*, 13(5): e0196391. `https://doi.org/10.1371/journal.pone.0196391`.