Emotion Recognition Using Audio Signal Processing

Arashdeep Mehroke Mohamed Hasan

October 20, 2024

1 Problem Statement

Human emotions play a critical role in social interactions, and recognizing these emotions from audio signals could significantly enhance human-computer interaction systems. The goal of this project is to develop a system that recognizes human emotions from audio signals by extracting relevant audio features from speech and applying machine learning models to classify emotions such as happiness, sadness, anger, etc.

We aim to leverage datasets like RAVDESS, CREMA, SAVEE, and TESS to train our models. These datasets contain a variety of emotional speech recordings, which will be used to create an accurate emotion classification system. The core challenge lies in processing audio data effectively and selecting the most relevant features, given that audio signals can be complex and noisy. Extracting meaningful features that accurately reflect emotional states, and training models to perform well across diverse emotional categories, is key to achieving robust results.

2 Proposed Methodology

Our approach to emotion recognition involves several steps:

2.1 Dataset Collection

We will utilize public datasets, including RAVDESS, CREMA, SAVEE, and TESS, each of which contains high-quality recordings of emotional speech. These datasets offer a variety of emotional categories, including happy, sad, angry, and neutral emotions, providing a comprehensive base for model training.

2.2 Feature Extraction

Key audio features such as Mel Frequency Cepstral Coefficients (MFCCs), Spectrograms, Zero-Crossing Rate, and Chroma features will be extracted using the Librosa library. These features represent both time-domain and frequency-domain aspects of the audio signals, capturing the nuances in emotional speech.

2.3 Dimensionality Reduction

To reduce the complexity of the feature space and enhance model performance, Principal Component Analysis (PCA) will be applied. PCA will help remove redundant features and retain only the most significant components, ensuring that the model trains efficiently without being affected by irrelevant data.

2.4 Model Selection

We will employ machine learning models, including but not limited to Support Vector Machines (SVM), Random Forest classifiers and DNN, for emotion recognition. These models have shown promise in previous studies due to their ability to handle complex, high-dimensional data. Hyperparameter tuning will be conducted using Grid Search to ensure optimal model performance.

2.5 Training and Testing

The models will be trained and tested using the selected datasets, split into training and testing subsets. Cross-validation will be used to assess the robustness of the models.

3 Evaluation Strategies

Model performance will be evaluated based on several metrics, including:

- **Accuracy:** The percentage of correctly classified emotions.
- **Precision:** The ratio of true positive predictions to the sum of true positive and false positive predictions.
- **Recall:** The ratio of true positive predictions to the sum of true positive and false negative predictions.
- **F1-Score:** The harmonic mean of precision and recall, providing a balance between the two.
- **Confusion Matrix:** A visualization of the classification performance, showing the distribution of true and predicted labels.

We plan to visualize these metrics using confusion matrices and performance curves, allowing us to assess how well the models differentiate between emotions.

4 Planned Evaluation

During training, we will monitor the accuracy, precision, recall, and F1-score of our models. In particular, we will evaluate how well each model performs across different emotion categories. A confusion matrix will help identify specific emotional classes where the models might be misclassifying, and we will use this information to adjust feature extraction or model tuning as necessary.

Additionally, hyperparameter tuning through Optuna will be employed to optimize the models. This tuning will focus on parameters including but not limited to the kernel type and regularization term for SVMs, the number of trees and maximum depth for Random Forest classifiers, and the number of layers and neurons for DNNs. By systematically exploring the hyperparameter space, we aim to improve model performance

5 Conclusion

This proposal outlines our methodology for developing a system that recognizes emotions from audio signals. By leveraging well-established audio features and advanced machine learning models, we aim to create an accurate and efficient emotion recognition system. Our planned evaluation strategies ensure that we can systematically improve model performance and gain insights into areas for further refinement.