Human Emotion Recognition from Audio Data

1. Introduction

Human emotions play a crucial role in communication and interpersonal relationships. Emotion recognition from audio data has gained significant attention due to its wide applications in areas such as healthcare, customer service, and human-computer interaction. This project focuses on building an emotion recognition system that can predict human emotions from audio recordings.

2. Dataset

We used the RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset, which contains audio samples labeled by emotion. The emotions include: Angry, Disgust, Fear, Happy, Neutral, Sad, Surprise. Each audio file is a `.wav` file containing an actor's emotional expression.

3. Feature Extraction

To effectively capture the relevant characteristics of audio signals, we performed feature extraction using the following techniques:
- MFCCs (Mel-Frequency Cepstral Coefficients): Captures timbral aspects of audio.
- Chroma STFT: Represents energy distribution across different pitch classes.
- Mel Spectrogram: Provides a representation of the short-term power spectrum of sound.
- Spectral Contrast: Measures the difference in amplitude between peaks and valleys in the spectrum.
- Tonnetz: Captures harmonic relationships in music and voice.

These features were extracted using the `librosa` Python library, and normalized using a standard scaler for model training.

4. Model Training

We used a Random Forest Classifier to train the model, chosen for its robustness and ability to handle high-dimensional feature spaces. The data was split into training and test sets in an 80:20 ratio. The model was trained using 500 estimators and a maximum depth of 50.

The final trained model, scaler, and label encoder were saved using joblib.

5. Streamlit Application

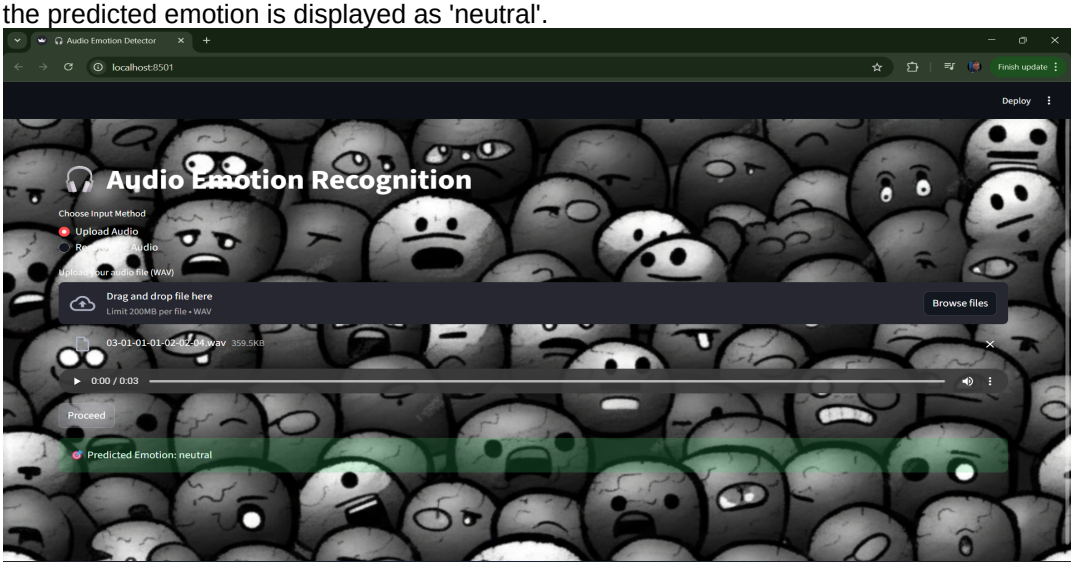
We developed a Streamlit-based web interface for user-friendly interaction with the model. The application allows two input methods:
- Upload Audio File (.wav)
- Record Live Audio (up to 20 seconds)

Once the user provides input and clicks 'Proceed,' the system extracts features, applies the scaler, and predicts the emotion using the trained model.

The interface includes a background image and displays the predicted emotion clearly.

6. Output Screenshot

The screenshot below shows the application interface where the user has uploaded an audio file, and the predicted emotion is displayed as 'neutral'.



7. Conclusion

This project successfully demonstrates real-time emotion recognition from audio files using feature extraction and a Random Forest Classifier. Future improvements could include using deep learning models such as CNNs or RNNs to further enhance accuracy and robustness.