Speech Emotion Detection Classifier Documentation

TABLE OF CONTENTS

5. **Data Visualisation and Exploration**
6. **Feature Extraction**
7. **Model Building**
- Data Splitting
- Data Augmentation
- Building the Convolutional Neural Network (CNN)
8. **Model Training**
9. **Model Evaluation**
- Confusion Matrix
- Classification Report
10. **Conclusion**
11. **References**t

1. Project Overview

Speech Emotion Recognition (SER) involves recognizing human emotions and affective states from speech signals. This project uses deep learning techniques to classify speech into different emotional categories. The ultimate goal is to apply SER in real-world scenarios, such as call centers and driver safety systems, to improve customer service and prevent accidents.

2. Data Sources

The project utilizes four different datasets:
- Ravdess: The RAVDESS dataset containing speech audio recordings labeled with various emotions.
- Crema: The CREMA-D dataset, a collection of audio files with emotions labeled as sad, angry, disgust, fear, happy, neutral.
- TESS: The Toronto emotional speech set (TESS) dataset with emotions categorized into sadness and surprise.
- Savee: The Surrey Audio-Visual Expressed Emotion (SAVEE) dataset, which contains audio clips with emotions such as angry, disgust, fear, happy, neutral, sad, and surprise.

3. Importing Libraries

In this section, the necessary Python libraries are imported to work with audio data, perform data analysis, and build deep learning models. These libraries include Pandas, NumPy, Librosa, Seaborn, Matplotlib, Scikit-Learn, and Keras.

4. Data Preparation

02 03 04

Data preparation involves creating dataframes for each of the four datasets (Ravdess, Crema, TESS, and Savee). This step includes organizing audio files by emotion and storing their file paths.

5. Data Visualisation and Exploration

This section provides data visualization, displaying the count of each emotion in the combined dataset using bar plots, allowing for a quick overview of the data distribution.

6. Feature Extraction

In speech emotion recognition, features are extracted from audio signals to train machine learning models. Common features include Mel-frequency cepstral coefficients (MFCCs) and Chroma feature extraction, which are essential for training deep learning models.

7. Model Building

This section outlines the process of splitting the data, augmenting the dataset to enhance model performance, and building a Convolutional Neural Network (CNN) for emotion recognition.

8. Model Training

The training process is described, involving the model being fitted to the training data to learn the relationships between audio features and emotions. Additionally, learning rate reduction and model checkpoint callbacks are used to improve training efficiency and save the best models.

9. Model Evaluation

Model evaluation is performed using confusion matrices and classification reports to assess the model's performance in classifying emotions accurately.

PROS AND Conse

Everest Cantu

Ceo Of Ingoude Company

Drew Holloway

Enhanced Customer Service

By using Speech Emotion Recognition (SER) in call centers, companies can better understand and categorize customer emotions, leading to improved customer service and issue resolutionc

Driver Safety

Implementing SER in car onboard systems can monitor the emotional state of the driver. It can help in preventing accidents by alerting the driver if they are in a distracted, agitated, or drowsy state.

Diverse Data Sources

The project utilizes multiple datasets (Ravdess, Crema, TESS, and Savee), which increases the diversity of emotional expressions and voices in the training data, making the model more robust

Deep Learning

The project employs deep learning techniques, such as Convolutional Neural Networks (CNNs), which are known for their ability to extract complex features from audio data and improve accuracy in emotion classification

Data Augmentation

Augmenting the dataset through techniques like pitch shifting and time-stretching can enhance model generalization and performance.

Cons

Biased Data

Datasets used for training may not be representative of the target user population, leading to potential bias in emotion recognition, especially if the training data does not include diverse voices or cultural backgrounds.

Data Preprocessing Challenges

Preprocessing audio data can be complex and time-consuming, involving feature extraction, scaling, and handling variations in audio quality, which can pose challenges.

Model Overfitting

Deep learning models can be prone to overfitting, where they perform well on the training data but poorly on unseen data. Regularization techniques and a sufficient amount of data are required to mitigate this issue

Computationally Intensive

Training deep learning models for SER can be computationally intensive and may require access to high-performance hardware or cloud resources.

Real-world Variability

Real-world scenarios may introduce additional challenges, such as background noise, overlapping speech, and varying emotional expressions, which the model may struggle to handle